A BAYESIAN SOLUTION FOR THE LAW OF CATEGORICAL JUDGMENT WITH CATEGORY BOUNDARY VARIABILITY AND EXAMINATION OF ROBUSTNESS TO MODEL VIOLATIONS

A Thesis
Presented to
The Academic Faculty

by

David R. King

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Psychology

Georgia Institute of Technology
December 2013

A BAYESIAN SOLUTION FOR THE LAW OF CATEGORICAL JUDGMENT WITH CATEGORY BOUNDARY VARIABILITY AND EXAMINATION OF ROBUSTNESS TO MODEL VIOLATIONS

Approved by:

Dr. James S. Roberts, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Christopher Hertzog
School of Psychology
Georgia Institute of Technology

Dr. Daniel H. Spieler
School of Psychology
Georgia Institute of Technology

Date Approved: November 1, 2013

ACKNOWLEDGEMENTS

I am very thankful for my advisor Jim Roberts' guidance throughout this project. His conversations with me about the topic and detailed feedback on drafts contributed greatly to the quality of the final product. I also want to thank my committee members, Christopher Hertzog and Daniel Spieler, for their insightful comments, and fellow student Zane Blanton, for his work on an earlier part of the project. Lastly, I am very thankful for the love and support of my beautiful wife, Jamie King.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Previous solutions for the the Law of Categorical Judgment with category boundary variability have either constrained the standard deviations of the category boundaries in some way or have violated the assumptions of the scaling model. In the current work, a fully Bayesian Markov chain Monte Carlo solution for the Law of Categorical Judgment is given that estimates all model parameters (i.e. scale values, category boundaries, and the associated standard deviations).

The importance of measuring category boundary standard deviations is discussed in the context of previous research in signal detection theory, which gives evidence of interindividual variability in how respondents perceive category boundaries and even intraindividual variability in how a respondent perceives category boundaries across trials. Although the measurement of category boundary standard deviations appears to be important for describing the way respondents perceive category boundaries on the latent scale, the inclusion of category boundary standard deviations in the scaling model exposes an inconsistency between the model and the rating method. Namely, with category boundary variability, the scaling model suggests that a respondent could experience disordinal category boundaries on a given trial. However, the idea that a respondent actually experiences disordinal category boundaries seems unlikely.

The discrepancy between the assumptions of the scaling model and the way responses are made at the individual level indicates that the assumptions of the model will likely not be met. Therefore, the current work examined how well model parameters

could be estimated when the assumptions of the model were violated in various ways as a consequence of disordinal category boundary perceptions.

A parameter recovery study examined the effect of model violations on estimation accuracy by comparing estimates obtained from three response processes that violated the assumptions of the model with estimates obtained from a novel response process that did not violate the assumptions of the model. Results suggest all parameters in the Law of Categorical Judgment can be estimated reasonably well when these particular model violations occur, albeit to a lesser degree of accuracy than when the assumptions of the model are met.

CHAPTER 1

INTRODUCTION

Louis L. Thurstone's (1927a, 1927b) scaling model is foundational to modern psychometric models of choice behavior (Bockenholt, 2006), as well as to models from several other areas of research, including the signal detection theory model in psychophysics (McNicol, 1972) and the random utility model in economics (McFadden, 2001). Thurstone's model maps psychological phenomena, such as attitudes, opinions, moral values, consumer preferences, utility, or aesthetic values, onto a latent continuum, where the stimuli are assigned quantitative values according to relative dominance. Once the scale is constructed, it can be used to measure individual differences; although the current work is focused solely on scale construction.

Scale construction begins with the development of stimuli. Following stimulus development, respondents judge the relative dominance of each stimulus according to one of several methods given by Thurstone. The current work examines the method of successive intervals (Saffir, 1937), a method in which respondents rate stimuli into $G+1$ graded response categories according to the degree to which each stimulus represents or contains the feature of interest. The aggregated ratings are recorded in a response matrix, accumulated across categories, converted to proportions, and finally, converted to z-scores. The z-scores can then be used to solve for the parameters of the latent scale through a series of equations called the Law of Categorical Judgment (LCatJ; Torgerson, 1958).

The parameters in the least constrained form of the LCatJ include stimulus scale values and associated standard deviations, as well as category boundaries and associated standard deviations. Although several solutions have been offered for the LCatJ (e.g., Torgerson, 1958; Bock & Jones, 1968; Rosner & Kochanski, 2009), no solution has estimated all four parameter types without violating model assumptions, and most solutions simply constrain the category boundary standard deviations to be equal.

Constraining the category boundary standard deviations simplifies the parameter estimation process, but the constraints may take away from the explanatory power of the model. Researchers using signal detection theory models, models that depend on many of the same assumptions as Thurstonian models, have investigated the way respondents perceive category boundaries. The results suggest there is interindividual variability in how respondents perceive category boundaries and even intraindividual variability in how a respondent perceives category boundaries across trials (see Benjamin, Diaz, & Wee, 2009 for a thorough review).

Although variability in perceptions of category boundaries may be an important characteristic of the latent scale, the inclusion of category boundary standard deviations in the scaling model exposes an inconsistency between the model and the acquisition of ratings. Namely, with category boundary variability, a respondent may experience disordinal category boundaries on a particular trial even though the true boundaries are ordinal. This can occur because the variability in the perception of category boundaries gives rise to category boundary distributions that will often overlap.

To examine the discrepancy between theory and practice when disordinal category boundaries are experienced, various hypothetical response processes[1] have been examined in the literature (Rosner & Kochanski, 2009; Klauer & Kellen, 2012). Response processes can be conceptualized as unconscious processes that allow a respondent to systematically rate a stimulus even when disordinal category boundaries are experienced[2]. There is no evidence that respondents use the response processes that have been examined, and the idea that respondents consciously experience disordinal category boundaries seems unlikely. However, investigations of hypothetical response processes offer a means to examine the extent to which the discrepencies between the scaling model and the acquired ratings affect the model parameter estimates.

The primary focus of the current work is to estimate category boundary standard deviations along with the other parameters commonly estimated in the LCatJ model. Because of the discrepancy between the scaling model and ratings made at the individual response level, the assumptions of the scaling model for the LCatJ will likely not be met. The current work will examine how well the parameters can be estimated when the assumptions of the model are violated in various ways. The following sections review Thurstone's scaling model, the derivation of the LCatJ, the evidence from the signal detection theory literature that suggests the importance of estimating category boundary standard deviations, and the extent to which the assumptions of the model are violated when various response processes are used.

---

[1] Response processes are called decision rules in the Rosner and Kochanski (2009) and Klauer and Kellen (2012) articles.

[2] The current work will not explore the theoretical mechanisms that may or may not give rise to various response processes.

Thurstone's Scaling Model

Thurstone (1927a, 1927b) theorized that a psychological stimulus elicits a different reaction from a respondent each time the respondent experiences the stimulus. The reaction, referred to as a discriminal process, can be thought of as the projected value of the stimulus on the psychological continuum at the moment the stimulus is experienced. A number of factors, including noise in the environment and the attentiveness of the respondent, cause the stimulus to be experienced in a slightly different way on each trial. The discriminal process for a stimulus on a particular trial is assumed to be randomly drawn from a normal distribution, which the stimulus projects on the latent scale. The mean of the discriminal process distribution is the scale value of the stimulus and the standard deviation is the amount of variability in the momentary values of the stimulus projections.

The response function given by Thurstone (1927a) for solving the scale values and standard deviations of the stimuli is called The Law of Comparative Judgment (LCompJ). The LCompJ relates the parameters of the scaling model to a matrix of aggregated stimulus judgments that are made according to the method of paired comparisons. Stimulus judgments are determined by the following theoretical process. When a respondent is presented with two stimuli, *j* and *k*, he or she unconsciously "perceives" two discriminal processes, d_j and d_k, each randomly drawn from its respective discriminal process distribution. The respondent then judges the stimulus with the higher discriminal process as the more dominant stimulus.

Each stimulus-stimulus pair is judged for a total of $J(J-1)/2$ judgments, where J is the total number of stimuli. Judgments are recorded in a $J \times J$ response matrix. The

response matrix records the number of times stimulus *j* is judged to dominate stimulus *k*, with *j*, *k* = 1 to *J* and *j* < *k*. The response matrix is complete after the same participant makes a large number of judgments across repeated trials for each stimulus-stimulus pair. The frequency of judgments in each cell of the response matrix is then divided by the total number of judgments made for each paired comparison and the proportion is converted to a z-score through a probit transformation. The full model for the LCompJ is defined as

$$x_{jk} = \left(s_k - s_j\right)\left(\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_j\sigma_k\right)^{-1/2} \quad (1)$$

where

s_j, s_k denote the scale values of stimuli j and k (i.e., the means of discriminal process distributions for d_j and d_k);

σ_j, σ_k denote the discriminal dispersions of stimuli j and k (i.e. the standard deviations of discriminal process distributions for d_j and d_k);

r_{jk} is the correlation between the pairs of discriminal processes d_j and d_k; and

x_{jk} is the normal deviate corresponding to the theoretical proportion of times stimulus k is judged greater than stimulus j.

The full model for the LCompJ is referred to as Case I when the same respondent repeatedly judges each stimulus pair or as Case II when each respondent in a group of respondents judges each stimulus pair once. The full model must be constrained in order to solve for the parameters. The simplest way to constrain the model is to make the assumption that the correlations between pairs of discriminal processes are zero. This is the least constrained version of the LCompJ that is identifiable. It is referred to as Case III and defined as

$$x_{jk} = \left(s_k - s_j\right)\left(\sigma_j^2 + \sigma_k^2\right)^{-1/2} \quad (2)$$

Although Thurstone proposed the LCompJ-Case III in 1927, solutions were not presented in the literature until years later. Advances in statistical analysis have led to a maximum likelihood solution using nonlinear optimization (Mackay & Chaiy, 1982) and a Bayesian solution using the Gibbs sampler (Yao & Bockenholt, 1999).

The Law of Categorical Judgment

The LCatJ was given by Torgerson (1958) as a generalization of Thurstone's LCompJ for use with graded response data collected according to the method of successive intervals. The LCatJ is identical to the LCompJ with the exception that stimulus *k* has been replaced with category boundary *g*. The full model of the LCatJ is defined as

$$x_{jg} = \left(t_g - s_j\right)\left(\sigma_j^2 + \sigma_g^2 - 2r_{jg}\sigma_j\sigma_g\right)^{-1/2} \tag{3}$$

where

t_g denotes the category boundary between response categories g and $g+1$;

s_j denotes the scale value of stimulus j;

σ_j denotes the standard deviation of stimulus j;

σ_g denotes the standard deviation of category boundary g;

r_{jg} is the correlation between discriminal processes for stimulus j and category boundary g; and

x_{jg} is the normal deviate corresponding to the theoretical proportion of times stimulus j is rated below boundary g.

By replacing stimulus *k* from the LCompJ with category boundary *g* in the LCatJ, Torgerson makes the assumption that category boundaries behave in the same way as stimuli. In other words, there are discriminal process distributions associated with each category boundary. Similar to the identification problem for the LCompJ, the LCatJ needs to be constrained in order to solve for the parameters. Setting the correlations between momentary values of stimulus *j* and category boundary *g* to zero eliminates the covariance terms from the model and the least constrained identifiable model for the LCatJ becomes

$$x_{jg} = \left(t_g - s_j\right)\left(\sigma_j^2 + \sigma_g^2\right)^{-1/2} \tag{4}$$

Equation 4 can be referred to as the LCatJ-Case III because of its similarity to the LCompJ-Case III. The identifiability of the model parameters is theoretically possible when:

$$2(J + G) \leq JG + 2 \tag{5}$$

where

J is the number of stimuli; and

G is the number of category boundaries.

The left side of the inequality represents the number of unknown values, whereas the right side of the inequality represents the number of known values. For the known values, the *JG* normal deviates are produced through the probit transformation of the cumulative proportions in the response matrix, and two of the scaling parameters can be fixed arbitrarily to give the scale a unit and an origin. The inequality is true when the scaling model has at least three category boundaries and at least four stimuli or at least four category boundaries and at least three stimuli.

The primary difference between the LCompJ and the LCatJ is the way the response matrix of stimulus judgments is formed. The LCompJ uses response data collected according to the method of paired comparisons, whereas the LCatJ uses response data collected according to the method of successive intervals. Scale values obtained from the method of successive intervals have been shown to be linearly related to scale values obtained from the method of paired comparisons (Saffir, 1937; Edwards & Thurstone, 1952).

The method of successive intervals was first presented by Thurstone's student Saffir (1937) as an alternative to the method of paired comparisons. Although the method of paired comparisons requires a participant to make $J*(J-1)/2$ judgments, the method of successive intervals requires a participant to make only J judgments, making the method of successive intervals a more practical method for data collection as J increases. For the original implementation of the method of successive intervals, respondents were asked to sort a set of stimuli into a predetermined number of piles according to the perceived dominance of the stimuli. The stimuli with the lowest perceived dominance were sorted into the first pile, and stimuli with higher amounts of perceived dominance were sorted into subsequent piles. Participants were asked to arrange the piles in such a way that the difference in perceived dominance between any two successive piles was equal.

In the modern implementation of the method of successive intervals, stimuli are rated using $G+1$ graded response categories ranging from $g = 1$ to $G+1$. According to the theory, when a respondent rates stimulus j in categories 1 through G, the respondent is judging stimulus j to be located at or below the gth category boundary, as well as at or

below each subsequent category boundary above the *g*th category boundary. In the case when a respondent rates stimulus *j* in response category *G+1,* the respondent is judging stimulus *j* to be located above the *G*th category boundary. Ratings are recorded in a *J* x (*G+1)* response frequency matrix. Each cell of the response frequency matrix contains the frequency of times that stimulus *j* is rated in response category *g*. If *N* respondents rate each stimulus once, the sum across each row of the response frequency matrix is equal to *N*. An example of a response frequency matrix created from simulated ratings is shown in Table 1a.

Although only one rating is made for each stimulus, the rating is used to derive *G* stimulus judgments. This is the number of judgments necessary to determine whether or not stimulus *j* is located at or below each of the *G* category boundaries. According to theory, the *G* stimulus judgments are derived by assuming that category boundaries maintain a rank ordering on the latent scale on each trial. This assumption allows for each stimulus judgment to be derived because a respondent rating stimulus *j* in response category *g* can be assumed to be judging stimulus *j* at or below the *g*th category boundary, as well as at or below each category boundary above the *g*th category boundary. In practical terms, the implicit stimulus judgments are derived by accumulating the frequencies in the response frequency matrix across categories, from *g = 1* to *G+1*, creating a cumulative response frequency matrix (Table 1b).

The cumulative response frequency matrix can be described as an implicit stimulus judgments matrix after the last column of the cumulative response frequency matrix is dropped because stimulus judgments are derived from the ratings, rather than explicity made by the respondent. For stimulus *j*, the cumulative frequency in category *g*

can be conceptualized as representing the number of times stimulus *j* is implicitly judged to be located below category boundary *g*. Response category *G+1* is not interpreted because there is not a category boundary above response category *G+1* and the cumulative frequency is equal to the number of times stimulus *j* is rated.

Table 1

Transformation of category response frequencies into z-scores

(a) *Response frequency matrix*

	c_1	c_2	c_3	c_4	c_5
s_1	138	149	131	75	7
s_2	42	107	168	144	39
s_3	17	56	114	169	144

Note. s_k = stimulus *k*; and c_g = response category *g*.

(b) *Cumulative response frequency matrix*

	c_1	c_2	c_3	c_4	c_5
s_1	138	287	418	493	500
s_2	42	149	317	461	500
s_3	17	73	187	356	500

Note. s_k = stimulus *k*; and c_g = response category *g*.

(c) *Cumulative proportions matrix*

	c_1	c_2	c_3	c_4	c_5
s_1	0.28	0.57	0.84	0.99	1.00
s_2	0.08	0.30	0.63	0.92	1.00
s_3	0.03	0.15	0.37	0.71	1.00

Note. s_k = stimulus *k*; and c_g = response category *g*.

(d) *Z-score matrix*

	t_1	t_2	t_3	t_4
s_1	-0.60	0.19	0.98	2.12
s_2	-1.38	-0.53	0.34	1.42
s_3	-1.83	-1.05	-0.32	0.56

Note. s_k = stimulus *k*; and t_g = category boundary *g*.

In order to convert the frequencies in the cumulative response frequency matrix into values that can be used in the LCatJ to solve for model parameters, the frequencies must first be converted into proportions and then converted into z-scores through a probit

transformation. The cumulative proportions matrix (Table 1c) is created by dividing the frequencies in each cell of the cumulative response frequency matrix by the total number of times the stimulus was rated. Finally, the probit transformation converts the proportions into z-scores, which indicate the standardized distance between each stimulus-category boundary pair on the latent scale. An example of this z-score matrix is shown in Table 1d.

The Link between Thurstonian Scaling and Signal Detection Theory

Before reviewing the evidence from the signal detection theory literature that indicates the importance of measuring category boundary standard deviations, it seems appropriate to briefly discuss the relevance of signal detection theory to Thurstonian scaling. In his paper on the theory of recognition, Tanner (1956) writes the following about Thurstone: "had he recognized the existence of a noise distribution such as the one postulated in the theories of detection and recognition, it seems likely that he would have developed essentially the same theory as that developed in the current set of papers [i.e. signal detection theory], only Thurstone would have been thirty years earlier" (p.888). The noise distribution Tanner is describing is the primary difference between the theory of Thurstonian scaling and signal detection theory. In Thurstonian scaling, stimuli are compared and ordered according to varying degrees of dominance. In signal detection theory, a stimulus is called a signal and is compared to noise.[3] The signal and noise

[3] Noise is defined as the absence of the signal.

distributions are assumed to be normally distributed, in the same way discriminal process distributions are assumed to be normal in Thurstone's scaling model.

The strength of the signal is measured by the probability that the signal is detected when the signal is present and not detected when the signal is absent. According to the theory, when the signal and noise distributions overlap, the probability of detecting the signal approaches chance (i.e. 0.5) as the amount of overlap increases. This mirrors the relationship between the probability that one stimulus is judged more dominant than another stimulus in the method of paired comparisons and the extent to which the associated discriminal process distributions overlap on the latent scale.

There are many concepts in Thurstonian scaling and signal detection theory that run parallel (even the receiver-operating characteristic in signal detection theory can be derived from the method of successive intervals; Lee, 1969). In signal detection theory, the method analogous to paired comparisons is called the forced choice task and the method analogous to successive intervals is called the rating scale task (McNicol, 1972). The link between Thurstonian scaling and signal detection theory is important because research in either area advances both theories.

The Importance of Measuring Category Boundary Variability

Signal detection theorists have gone to a great deal of effort over the last several years to test whether the equal category boundary[4] variance model or the unequal category boundary variance model is better suited for modeling detection scenarios.

---

[4] A category boundary is called a criterion in the signal detection literature.

Benjamin, Diaz, and Wee (2009) present several reasons to suggest that there is variability in category boundary perceptions.

First, category boundary values do not remain constant from study to study or even within studies. The authors hypothesize that factors such as hunger, boredom, or distractions from the surrounding environment could account for changes in the perception of category boundary values over the course of an experiment. Further, autocorrelations of responses across trials, even when stimulus scale values are independently drawn, indicate that perceptions of category boundaries depend on previous perceptions of category boundaries, namely the perceptions of category boundaries from the most recent trials.

A second reason Benjamin, Diaz, and Wee give for category boundary variability is the demand that is placed on working memory when a participant is required to remember several category boundaries throughout an experiment. The authors point to studies that show respondents are better able to discriminate between new and previously learned words when fewer categories are used in the rating task.

A third reason for category boundary variability given by the authors is that Ratcliff's diffusion model[5] requires category boundary values to be changed throughout an experiment in order to fully account for various response times and accuracies. Finally, the authors discuss experiments in which the researchers manipulated the amount of attention participants gave to a task. Results indicated that category boundary variances were smaller when participants devoted more attention to the task. Following

---

[5] For a review of the variations of Ratcliff's diffusion model see Ratcliff and Smith (2004).

the literature review, Benjamin, Diaz, and Wee proposed a new theory called the Noisy Decision Theory of Signal Detection, which parameterizes category boundary variability for the purpose of explaining response data from several detection tasks that are not fully explained by the traditional signal detection theory model.

The Discrepancy between the Scaling Model and the Rating Method

Considering the evidence of variability in category boundary perceptions for signal detection tasks, it seems reasonable to assume that respondents rating stimuli according to the method of successive intervals would also show variability in category boundary perceptions. Although this variability may be important to model, a discrepancy between the scaling model and the rating method makes the task of estimating category boundary standard deviations more difficult.

The discrepancy arises because under the assumption that category boundaries project normal distributions on a latent scale, respondents may experience disordinal category boundaries on a given trial when the category boundary distributions overlap. If Thurstone's scaling model is correct and ratings are based on momentary values from the stimulus and category boundary discriminal process distributions, then the possibility exists that a momentary value of stimulus j is both lower than the momentary value of category boundary g and higher than the momentary value of category boundary $g+1$. If a respondent really does experience discriminal processes according to the assumptions of the scaling model, then it is not clear how the respondent rates the stimulus when disordinal category boundaries are experienced.

Realistically, it does not seem possible that a respondent could experience disordinal category boundaries on a given trial, because the respondent simply views

category boundaries as ordered integers. The respondent may experience category boundaries as having different values on the latent scale for each trial, but for any given trial, the respondent will pragmatically experience the category boundaries as ordered. This makes the experience of category boundaries distinctly different than the experience of stimuli. Category boundaries have an implicitly predetermined rank order, whereas stimuli do not. This discrepancy between the scaling model and the rating method may suggest that category boundaries do not project normal distributions on the latent scale.

To illustrate the problem, a graphical representation of the way respondents are theorized to experience discriminal processes according to the assumptions of the scaling model is shown in Figure 1. The arrows represent momentary values of the category boundaries experienced by a respondent on a particular trial. Although the means of the discriminal process distributions indicate the category boundaries are located in sequential order, the momentary values are in a different order than the means. Namely, according to theory, the respondent experiences category boundary two to be below category boundary one. If the respondent experiences stimulus *j* to be in the region of the scale between momentary values of category boundary one and category boundary two, then the respondent can either rate the stimulus above category boundary two (e.g., response category three) or below category boundary one (e.g., response category one). The various ways stimuli can be rated when disordinal category boundaries are experienced leads to different relative frequencies across categories in the response matrix, and ultimately different parameter estimates.

It is not clear how the scaling model can be reconciled with the rating method. However, the way judgments between stimuli and category boundaries should be made

according to the scaling model, and the way judgments are actually made according to the method of successive intervals can be examined.

The Consistent Response Process

When the method of successive intervals was discussed earlier, it was touted for requiring fewer responses than the method of paired comparisons requires. Fewer responses are required for the method of successive intervals because judgments are derived from ratings rather than explicitly made. Namely, the assumption that category boundaries are ordered allows *G* judgments, one for each category boundary, to be derived from a single rating. The *G* judgments are derived by accumulating the ratings in the response matrix (Table 1a) across categories, from $g = 1$ to $G+1$. The resulting matrix is the cumulative response frequency matrix (Table 1b).

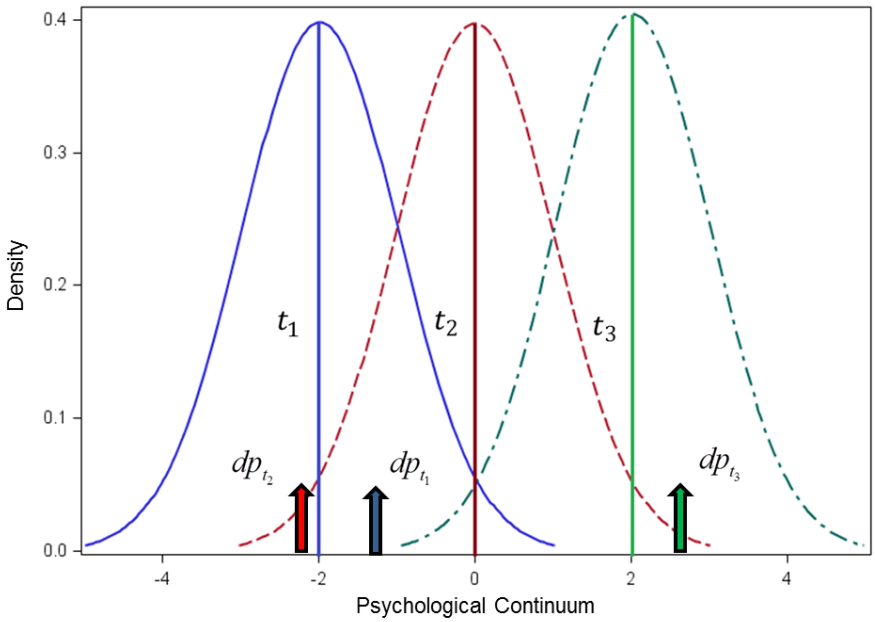


Figure 1. Graphical representation of disordinal category boundaries experienced by a respondent on a particular trial. *Note*: t_g = location of category boundary g on the psychological continuum; dp_{tg} = discriminal process for category boundary t_g.

According to Thurstone's scaling model, respondents may not experience ordinal category boundaries on a given trial if discriminal process distributions for category boundaries overlap. In this case, the assumption necessary for accumulating the frequencies in the response matrix may not hold. If the respondent does not experience ordered category boundaries on a given trial, the implicit judgments derived from the rating will not be consistent with the explicit judgments that would have been made if stimulus *j* had been judged against each category boundary separately.

The method of successive intervals can be altered so that explicit judgments are obtained. The method requires a respondent to judge whether or not each stimulus is located below each category boundary. The response process used for this method will be referred to throughout the paper as the *Consistent Response Process*, and can be stated as follows: *for a given stimulus* s_j*, calculate the G differences* $t_g - s_j$*. For every positive difference, make response g*. The *G+1*th category is not used because there is not a category boundary above the *G+1*th category. This does not affect the transformation of response frequencies into z-scores because the *G+1*th category is dropped in the transformation process.

The matrix obtained from using the Consistent Response Process is similar to the cumulative response matrix obtained in the traditional approach. The reason the matrices are similar is because, for the Consistent Response Process, responses do not need to be accumulated across categories. Each judgment is explicitly made and consequently, there are no judgments that need to be derived from accumulating responses across categories. In obtaining the z-scores, the response frequencies from the Consistent Response Process can simply be converted into proportions (Table 1c) and then z-scores (Table 1d). This is

analogous to starting the transformation process with the cumulative response frequency matrix (Table 1b) instead of the response frequency matrix (Table 1a).

It is important to note that the matrix obtained from using the Consistent Response Process is not truly a cumulative response matrix. This is because the frequencies in each cell are not constrained to be equal to the sum of the frequencies in the preceding cells. Because the frequencies are not cumulative, there is a possibility that for a given stimulus, the response frequency in category $g+1$ could be smaller than the response frequency in category g. If this occurred, then subtracting adjacent categories would yield negative response frequencies. The probability of categories having nonmonotonic frequencies from $g = 1$ to $G+1$ will decrease as the amount of overlap decreases between category boundary discriminal process distributions and as the number of respondents increases. (A pilot study described later in this proposal suggests that, the possibility of negative response frequencies occurring is very small for 500 respondents.)

Although the Consistent Response Process requires the respondent to make a greater number of responses than are typically made for the method of successive intervals, the advantage of the Consistent Response Process is that the judgments are consistent with discriminal process values, even when disordinal category boundaries are experienced. In practice, respondents most likely do not experience category boundaries as disordinal, and thus, the Consistent Response Process may not provide any real-world data collection advantages. However, the Consistent Response Process can be useful in data simulation studies when the true parameters are known. Namely, the process can be used to create an explicit judgment matrix that is similar to the cumulative response frequency matrix and consistent with the assumptions of the LCatJ model.

Inconsistent Response Processes

Researchers have been aware of the problem of disordinal category boundaries for decades (e.g., Sjoberg, 1964). However, recently there have been efforts to better understand how the LCatJ works if the scaling model and rating process are inconsistent. Rosner and Kochanski (2009) showed that category boundary variability can lead to theoretical cumulative probabilities at category *g* that are greater than the theoretical cumulative probabilities at category *g+1*. These nonmonotonic cumulative probabilities may yield negative theoretical probabilities for response *g+1* when such probabilities are derived by differencing adjacent cumulative probabilities. This finding is in line with the discussion in the previous section about how negative response frequencies can be obtained when using the Consistent Response Process. To preclude the prediction of negative response probabilities, Rosner and Kochanski developed a probability density function for the LCatJ they termed "The Law of Categorical Judgment: Corrected." They claimed the probability density function correctly models the independence of category boundary discriminal processes.

To explain how ratings can be determined even when category boundaries overlap, Rosner and Kochanski gave a response process that will be referred to throughout the current paper as *Response Process 1*: *for a given stimulus* s_j*, calculate the G differences* $t_g - s_j$*. If* $t_g - s_j$ *is the smallest positive difference, then make response g. If all* $t_g - s_j$ *are negative, then make response G+1.*

In response to the "Law of Categorical Judgment: Corrected," Klauer and Kellen (2012) made the point that a number of response processes could be used to correct the LCatJ and the particular response process selected could lead to different response

distributions. The authors showed that the response process offered by Rosner and Kochanski produces response distributions with asymmetric probabilities (i.e. relative frequency distributions that are skewed rather than symmetrical). Klauer and Kellen then offered a response process that produces response distributions with symmetric probabilities, which will be referred to throughout the current paper as *Response Process 2*: *for a given stimulus s_j, calculate the G absolute differences $|t_g - s_j|$. If $|t_g - s_j|$ is the smallest absolute difference and $t_g - s_j$ is positive, then make response g. If $|t_g - s_j|$ is the smallest absolute difference and $t_g - s_j$ is negative, then make response g+1.*

From the assertion Klauer and Kellen made that a number of response processes can be used to rate stimuli, the current work will examine a new response process, referred to throughout the paper as *Response Process 3*: *For a given stimulus s_j, calculate the G differences $t_g - s_j$. Of the $t_g - s_j$ that are positive, find the smallest g and make response g. If all $t_g - s_j$ are negative, then make response G+1.*

The Relationship between Response Processes and Category Frequencies

Each response process relies on a different rule for classifying the stimulus into one of the *G+1* response categories. Subsequently, the particular response process the respondent uses will have some effect on the category response probabilities for a given stimulus. This raises an important question: How much do the response probabilities of a stimulus vary depending on the response process used to rate the stimulus? To answer this question, data were generated from discriminal process distributions and rated according to each of the four response processes (i.e., the Consistent Response Process, Response Processes 1-3). The category boundary discriminal process distributions were

equally spaced with mean values specified as integers one through eight. The standard deviations of the discriminal process distributions of stimuli and category boundaries were fixed at unity. Finally, two stimuli, one with a scale value in the center of the latent scale and one with a scale value to the left of the midpoint of the scale were specified. Each stimulus was systematically rated 100,000 times using each of the four response processes.

Figure 2 shows the relative frequencies of the category responses for the four response processes used to rate the stimulus located in the center of the scale. The Consistent Response Process appears to be normally distributed, but this is not surprising because the ratings are based on discriminal process difference scores that are normally distributed. For the Consistent Response Process, there is not any data lost in the rating process because all ratings are explicitly made between category boundaries and the stimulus. This allows the response probabilities to maintain normality.

The relative frequencies obtained from using Response Processes 1-3 were surprisingly similar to the relative frequencies obtained from using the Consistent Response Process. As Klauer and Kellen mentioned, the relative frequency distribution for Response Process 1 is negatively skewed, but only slightly, and the relative frequency distribution for Response Process 3 is slightly skewed as well, but in the positive direction. Finally, the relative frequencies look similar for Response Process 2 and the Consistent Response Process.

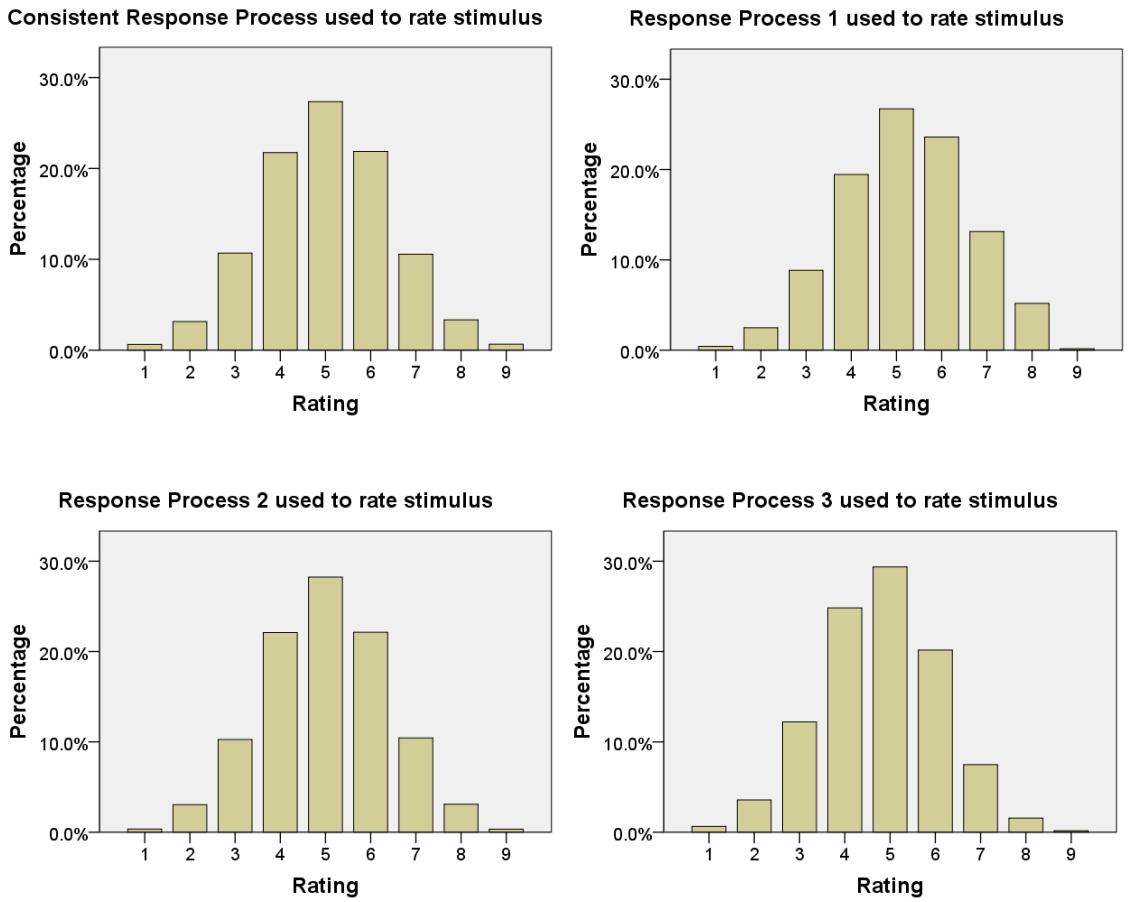


Figure 2. Relative frequencies of category responses based on the response process used to rate the stimulus. The stimulus scale value was set at 4.5 and the standard deviation was fixed at unity. Top left = Consistent Response Process; top right = Response Process 1; bottom left = Response Process 2; bottom right = Response Process 3.

The second stimulus rated in the simulation study was assigned a scale value of 1.5, putting its mean far to the left of the midpoint of the scale. From looking at the relative frequency distributions in Figure 3, the effect of the response processes on the response probabilities appears to be much greater for the stimulus assigned a scale value of 1.5 than the effect of the response processes on the response probabilities for the stimulus assigned a scale value in the center of the scale. Of particular note is how the modal response changes depending on the particular response process used. Response Process 3 is the only response process that produced a distribution with the same modal

response as the Consistent Response Process. These relative frequency distributions may indicate that parameters on or near the outer sections of the scale may be more difficult to estimate accurately because the distributions show greater variability across response processes than the variability across response processes for the stimulus in the center of the scale.

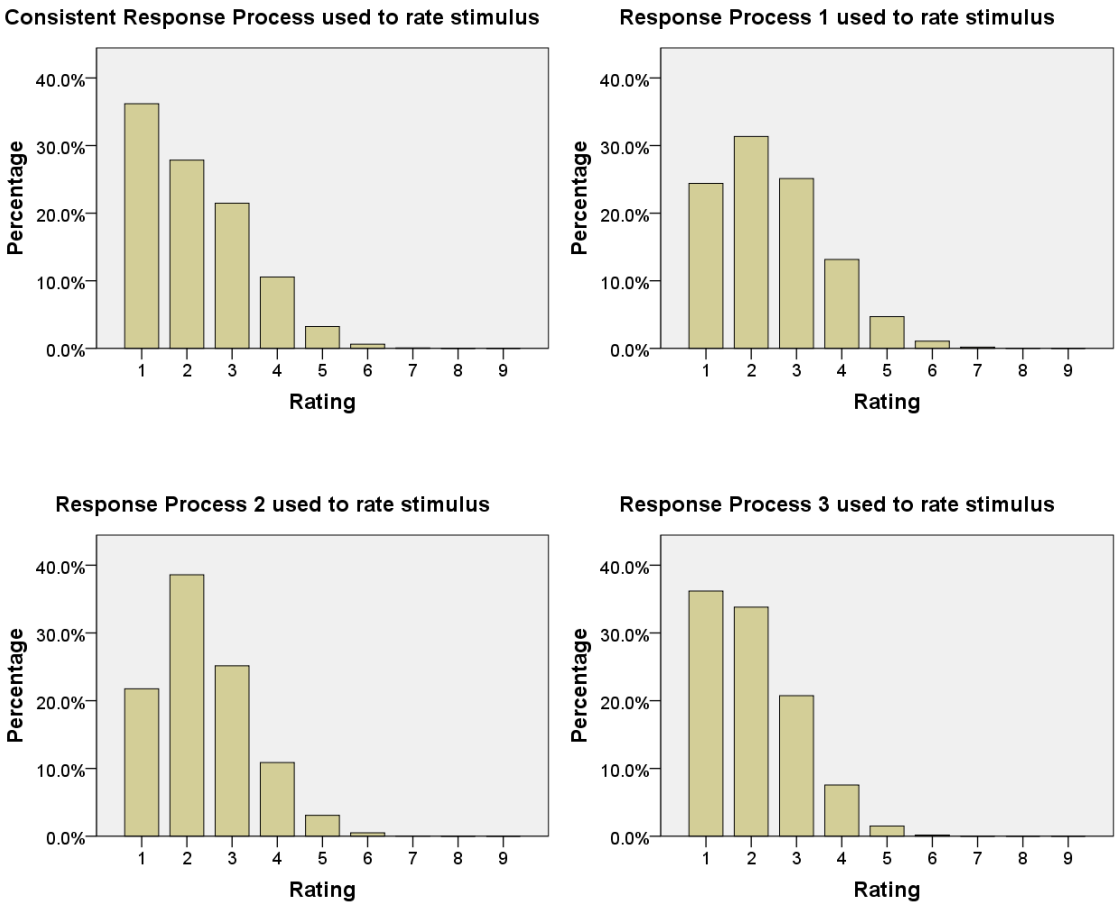


Figure 3. Relative frequencies of category responses based on the response process used to rate the stimulus. The stimulus scale value was set at 1.5 and the standard deviation was fixed at unity. Top left = Consistent Response Process; top right = Response Process 1; bottom left = Response Process 2; bottom right = Response Process 3.

The Consistent Response Process is in line with the assumption of normal discriminal process distributions because each judgment is explicitly made and none of

the independently drawn discriminal processes are ignored. However, Response Processes 1-3 are not in line with the assumption of normal discriminal process distributions because many of the judgments are implicitly made. The implicit judgments effectively ignore disordinal discriminal process samples because implicit judgments are made under the assumption that discriminal process samples are ordinal..The amount that the response distributions using Response Processes 1-3 deviate from the response distribution using the Consistent Response Process is an indication of the extent to which Response Processes 1-3 are not in line with the assumption of normal discriminal process distributions. Rather, the response distributions for Response Processes 1-3 are indicative of some other distribution of discriminal processes, thus suggesting a violation of the LCatJ assumptions.

Goals of the Current Study

Evidence from the signal detection theory literature suggests that the estimation of category boundary standard deviations allows the scaling model to more fully account for the response data collected from judgment tasks. Torgerson's least constrained model for the LCatJ [Equation 4] can be used to solve for these category boundary standard deviations along with the other parameters commonly estimated in the model.

Rosner and Kochanski (2009) presented a solution to Torgerson's least constrained model for the LCatJ that involved maximizing the log likelihood of the data to obtain model estimates and then sampling from a Bayesian posterior density to obtain confidence limits around the estimates. Their model estimates were strongly correlated with the true values, although they had difficulty recovering the correct scale. The loss of the scale may have resulted from the prior distributions used in the solution, which may

have pulled the estimates away from the true parameters. The loss of the scale may also have resulted from a scale constraint that was relatively weak in influence. In a follow-up article, Kochanski and Rosner (2010) used a set of stronger constraints, including the constraint that category boundaries must be in ascending order. Again, the scale of the estimates appeared to drift from that of the true parameters, which suggests that the constrained parameters were not responsible for the misestimation of the unit of the latent scale. Instead, this suggests that the misestimation of the true scale was primarily due to the fixed hyperparameters of the prior distributions pulling the estimates away from the true scale values. In the follow-up article, Kochanski and Rosner rescaled the generating parameters to match the scale of the recovered parameters. This was reasonable because the underlying scale was interval in nature and, therefore, the origin and unit of the scale were arbitrary. Scatterplots of the results indicated the estimates were accurate for all four parameter types.

However, the constraint that category boundaries must be in ascending order seems to ignore the problem of disordinal category boundaries discussed in the current paper. Further, the constraint violates the assumption of normally distributed category boundary distributions. If the momentary values of category boundaries are drawn sequentially, the first category boundary is drawn from a normal distribution, but the other category boundaries are not drawn from normal distributions because of the left censoring imposed by the constraint. The assumption of normally distributed discriminal process distributions will be violated to the extent that the category boundary distributions overlap.

Although a solution has been given for the least constrained version of the LCatJ (Kochanski & Rosner, 2010), the solution violates the assumption of normally distributed discriminal process distributions because of the ordinality constraint. Further, there appears to be ambiguity surrounding the way the model needs to be constrained for accurate estimates to be obtained. A primary goal of the current study is to present a fully Bayesian Markov chain Monte Carlo (MCMC) solution and to accurately recover the generating parameters. The current study explores the extent to which parameters can be accurately estimated when the category boundary discriminal distributions are independent and the order of the momentary values of the category boundaries is not restricted.

The assumptions for the original LCatJ will not hold in the successive intervals procedure when category boundaries are disordinal. Previous research has not explored the extent to which parameter estimates obtained from using a particular response process deviate from the parameter estimates that would have been obtained if the rating problem did not exist. Thus a second goal of the current study is to use the Consistent Response Process to make explicit stimulus-category boundary judgments and then compare the parameter estimates obtained from the MCMC solution to estimates obtained from response processes that violate the assumptions of the scaling model. Results will indicate the robustness of the estimates to violations of model assumptions.

Lastly, the current study examines the effects of the number of stimuli and the number of category boundaries on the accuracy of the estimates. Previous research (Jones, 1959) has shown stimulus scale values and standard deviations to be invariant to changes in the number of response categories and even to changes in the labels of the

categories. The current study examines whether this invariance holds with category boundary variability and the violation of model assumptions.

CHAPTER 2

METHOD

Experimental Design

Estimation accuracy was assessed through a parameter recovery simulation in which response process, number of stimuli, and number of response categories varied across replications. Response process consisted of four levels: the three response processes described earlier in the paper (i.e. Response Processes 1-3) and the response process devised to obtain judgments that are consistent with the original assumptions of the LCatJ (i.e. the Consistent Response Process). Number of stimuli contained two levels: ten stimuli and 20 stimuli. Number of categories contained two levels: five categories and nine categories. Thirty replications were conducted for each cell of the factorial design.

For each replication, model parameters were estimated using the fully Bayesian Markov chain Monte Carlo (MCMC) method. Average discrepancy between true and estimated values was indexed by root-mean-square deviation (*RMSD*) for each parameter type: scale values, stimulus standard deviations, category boundaries, and category boundary standard deviations.

$$\text{RMSD} = \left\{ \frac{\sum_{j=1}^{N} \left[\hat{\beta}_j - \beta_j \right]^2}{N} \right\}^{1/2} \quad (6)$$

where

$\hat{\beta}_j$ is the estimated value of the *j*th parameter of a given type of parameter;

β_j is the true value of the *j*th parameter of a given type of parameter; and

N is the number of parameters of a given type of parameter (e.g., 10 stimulus standard deviations).

Roberts and Laughlin (1996) showed that Equation 6 is equal to the following equation:

$$\text{RMSD} = \left\{ S_{\hat{\beta}}^2 + S_{\beta}^2 - 2S_{\hat{\beta}\beta} + \left[\bar{X}_{\hat{\beta}} - \bar{X}_{\beta} \right]^2 \right\}^{1/2} \quad (7)$$

where

$S_{\hat{\beta}}^2$ is the sample variance of the N $\hat{\beta}$s;

S_{β}^2 is the sample variance of the N βs;

$S_{\hat{\beta}\beta}$ is the sample covariance between $\hat{\beta}$ and β;

$\bar{X}_{\hat{\beta}}$ is the average of the N $\hat{\beta}$s; and

$\bar{X}_{\beta}$ is the average of the N βs.

This formulation suggests that three measures can be used to better explain the RMSD: the correlation between estimated and true values, the absolute mean difference between estimated and true values, and the ratio of variances between estimated and true values. Although RMSD served as the dependent measure in the experimental design, the other measures of mean accuracy were calculated to aid in the interpretation of model effects.

Data Generation

Data were generated using the SAS statistical software. Scale values were randomly sampled from uniform distributions with means equally spaced from negative one to eight for the nine category condition, and from negative one to four for the five category condition. Category boundaries were randomly sampled in the same manner as stimuli, although the means were equally spaced from zero to seven for the nine category condition, and from zero to three for the five category condition. These uniform distributions had a range of .5 and did not overlap. Consequently, true category boundaries maintained ordinality on the latent scale for each replication. Finally, stimulus and category boundary standard deviations were randomly sampled from uniform distributions with a mean of 1.25 and a range of 1.5.

All parameters were randomly sampled for each replication within a cell. A response matrix for each replication within a cell was created from the ratings of five hundred simulated participants. Each simulated participant rated each stimulus once. The process of converting the ratings from the response matrix to the z-score matrix follows the procedure illustrated in Table 1.

Markov Chain Monte Carlo Estimation

The parameters of equation 4 were estimated using a Markov chain Monte Carlo method. Parameter estimates were defined as the means[6] of the posterior distributions.

---

[6] Thurstone (1927a) originally defined scale values as the modes of their respective discriminal process distributions. However, this is an arbitrary distinction because the modal values will be the same as the mean values as long as the assumption of normally distributed discriminal process distributions holds.

For the prior distributions, normal distributions with means of .01 and precisions of .01 were used for sampling stimulus scale values and category boundaries. The normal distributions were ideal because the discriminal process distributions are normal in Thurstone's scaling model. However, the hyperparameters make the prior distributions very unrestrictive. For sampling stimulus and category boundary variances, inverse gamma distributions with shape and rate parameters of .5 were used. The inverse gamma is commonly used to model sample variances in MCMC (Gill, 2002) and these hyperparameters yield relatively loose prior distributions.

A burn-in period of 300,000 iterations was discarded before sampling from the posterior distribution. The number of iterations was chosen to ensure convergance and because of the low cost (i.e. high speed) of running a large number of iterations. A thinning factor of 200 was used to reduce the autocorrelations between successive draws from the chain. The chain ran for a total of 500,000 iterations. The 200,000 post burn-in iterations divided by the thinning factor of 200 gave an effective sample size of 1000 draws.

OpenBUGS, a Bayesian modeling program, was used to estimate the parameters. The 'coda' package (Plummer, Best, Cowles, & Vines, 2006) for *R* was used to calculate the convergence diagnostics.

CHAPTER 3

RESULTS

Convergence Assessment

Convergence of the Markov chain was first assessed through visual examination of the trace plots. For each condition in the experimental design, plots appeared to be stationary across the portion of the chain from which the sample was drawn. Prototypical trace plots for scale value, category boundary, stimulus standard deviation, and category boundary standard deviation posterior distributions are shown in Figures 7-10 of Appendix I.

Convergence was also assessed with statistical convergence diagnostic criteria. For Geweke's (1992) diagnostic, a separate z-value for each parameter in the chain was calculated. Each z-value was calculated as the difference between the sample means of the first 10% and the last 50% of the chain, divided by the estimated standard error. Across conditions, 82-92% of the z-values did not exceed an absolute value of two, providing evidence for the stationarity of the chain. For Heidelberger and Welch's (1983) diagnostics, a separate stationarity test and half-width test were calculated for each parameter in the chain. The stationarity test uses the Cramer-von-Mises statistic to test the null hypothesis that the sampled values from the chain come from a stationary distribution. If the null is not rejected for the stationarity test, a half-width test is conducted in which half the width of a 95% confidence interval around the sample mean is compared to a rough estimate of the variance of the sample mean. If the half-width is smaller than the estimated variance, then the test is passed and the chain is assumed to

have converged. Across conditions, 100% of the stationarity tests were passed and 81-91% of the half-width tests were passed. Summary results from the convergence tests are provided in Table 3 of Appendix I.

Parameter Recovery

Four separate univariate split-plot ANOVAs were conducted to examine the effect of response process, number of stimuli, and number of response categories on the average RMSD between the recovered parameters and the generating parameters for each parameter type (i.e. scale values, category boundaries, stimulus standard deviations, and category boundary standard deviations). Response process was treated as a within-replication factor because the initial data were identical across response process conditions prior to the rating of the stimuli. For each within-replication effect, the degrees of freedom of the mean square ratio were adjusted using the Huynh-Feldt correction.

The number of stimuli and the number of response categories served as between-replication factors in the split-plot analyses. Because four dependent measures were examined, the significance level was set at $\alpha = .05/4 = .0125$ when testing the effects. Between-replication and within-replication proportions of variation (η^2_{wf}) were calculated for each dependent measure. The symbol η^2_{wf} denotes the within-family proportion of variation accounted for by an effect. Only effects that were both statistically significant and accounted for at least 10% of the within-family variation in a dependent measure were interpreted. The effect size cutoff seemed reasonable for detecting the most important differences in estimation accuracy and has been used in similar parameter recovery studies (e.g., Roberts, Donoghue, & Laughlin, 2002). In hindsight, the cutoff

value may have been too small, as even effects that accounted for small differences in estimation accuracy were deemed worthy of interpretation.

Within-family eta squared values for all effects are shown in Table 2. Statistically significant effects are italicized and effects accounting for at least 10% of the within-replication proportion of variation in estimation accuracy are shown in bold. The total within-family sum of squares for each dependent measure is given in parentheses. All effects exceeding the η^2_{wf} cutoff were also statistically significant, essentially making the effect size cutoff the only criteria for the identification of important effects. The mean accuracy measures for each parameter type are shown in Appendix B.

Response process had a large effect on the accuracy of the recovered scale values, η^2_{wf} = .556, and category boundaries, η^2_{wf} = .788. Scale values were most accurately estimated when the Consistent Response Process was used to rate stimuli, *RMSD* = .216, and estimated to a lesser degree of accuracy when Response Process 3, *RMSD* = .248, Response Process 1, *RMSD* = .268, and Response Process 2, *RMSD* = .332, were used to rate stimuli. Category boundaries were also most accurately estimated by the Consistent Response Process, *RMSD* = .152, and estimated to a lesser degree of accuracy by Response Process 3, *RMSD* = .365, Response Process 1, *RMSD* = .494, and Response Process 2, *RMSD* = .689. A familywise error rate of .0125 was set for the post hoc analysis for each parameter type. Using the Bonferroni correction, six paired-samples t-tests were conducted at α = .0125/6 = .0021 for each parameter type. Mean accuracy estimates for scale values and category boundaries were statistically different between each pair of response processes.

Table 2

Eta-squared within family (η^2_{wf}) values for ANOVA effects on estimation accuracy by parameter type

Effect	RMSD			
	$\hat{s}_j$	$\hat{\sigma}_j$	$\hat{t}_g$	$\hat{\sigma}_g$
Within-replication				
RP	***0.556***	***0.101***	***0.788***	***0.642***
RP x S	*0.018*	***0.101***	*0.021*	*0.081*
RP x C	*0.034*	*0.067*	*0.083*	*0.053*
RP x S x C	0.003	0.011	0.001	*0.008*
Model	0.611	0.28	0.893	0.784
(SS_{wf})	(1.557)	(0.267)	(23.226)	(6.852)
Between-replication				
S	0.011	0.031	***0.253***	***0.382***
C	0.014	*0.082*	***0.314***	*0.046*
S x C	0.001	0.021	0.019	0.003
Model	0.026	**0.134**	**0.586**	**0.431**
(SS_{wf})	(0.972)	(1.59)	(8.356)	(9.609)

Note. SS_{wf} = sum of squres within family; RP = response process; S = stimuli; CB = category boundaries; $\hat{s}_j$ = scale values; $\hat{\sigma}_j$ = stimulus standard deviations; $\hat{t}_g$ = category boundaries; and $\hat{\sigma}_g$ = category boundary standard deviations. The η^2_{wf} for the model is the sum of the η^2_{wf} values across all experimental effects within a family. Statistically significant effects are italicized and effects with an η^2_{wf} value $\geq .1$ are shown in bold.

The average discrepancies of scale value estimates and category boundary estimates can be directly compared because the true scale values and true category boundaries had the same standard deviations. Although response process accounted for 79% and 56% of the within-family variation in estimation accuracy for category boundaries and scale values respectively, there was a much larger amount of variability to account for in category boundary accuracy (SS_{wf} = 23.226) than in scale value accuracy (SS_{wf} =1.557). Indeed there was approximately 15 times more variability in category boundary estimation accuracy than in scale value estimation accuracy. These differences are shown for each response process in Figure 4.

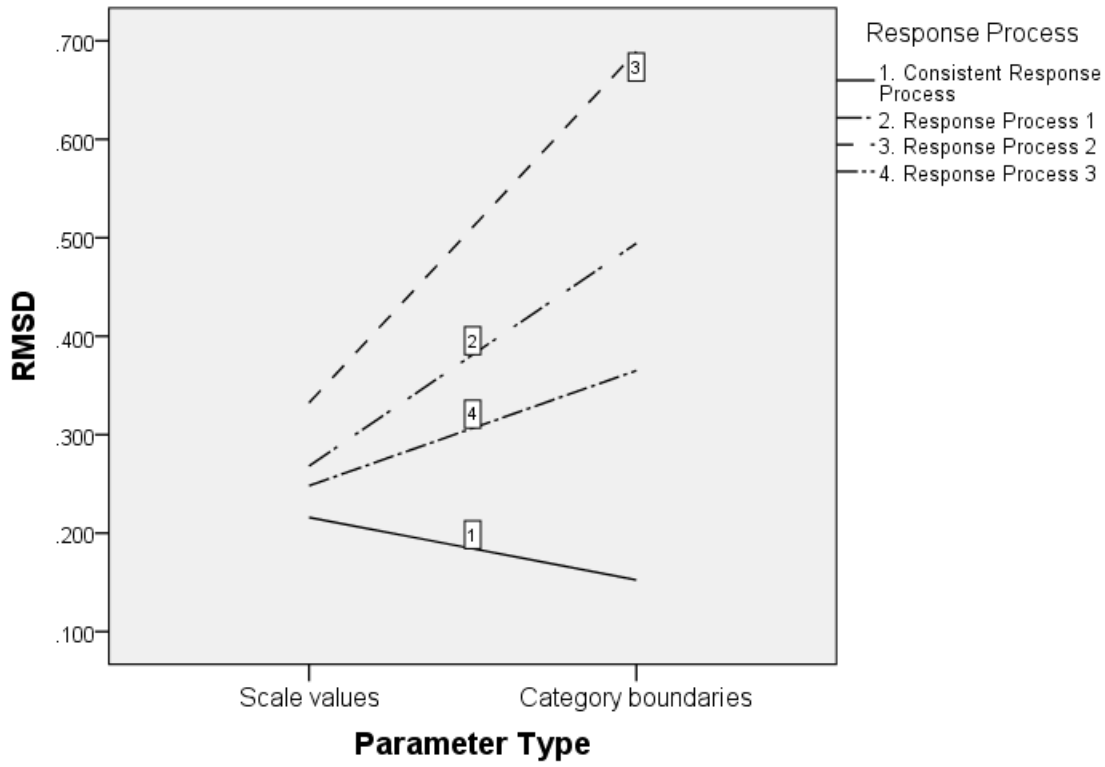


Figure 4. Root-mean-square deviations for scale values and category boundaries by response process.

Response process also had a large effect on the accuracy of the recovered category boundary standard deviations, η^2_{wf} = .642, and a small effect on the accuracy of the recovered stimulus standard deviations, η^2_{wf} = .101. Category boundary standard deviations were most accurately estimated by Response Process 3, *RMSD* = .373, and estimated to a lesser degree of accuracy by the Consistent Response Process, *RMSD* = .401, Response Process 1, *RMSD* = .539, and Response Process 2, *RMSD* = .604. Stimulus standard deviations were also most accurately estimated by Response Process 3, *RMSD* = .336, and estimated to a lesser degree of accuracy by Response Process 2, *RMSD* = .349, Response Process 1, *RMSD* = .352, and the Consistent Response Process, *RMSD* = .356.

A post hoc analysis was conducted using a familywise error rate of .0125 for each parameter type. Using the Bonferroni correction, six paired-samples t-tests were conducted at α = (.0125 / 6) = .0021 for each parameter type. For the recovered category boundary standard deviations, the estimation accuracy was not statistically different between Response Process 3 and the Consistent Response Process. Further, category boundary standard deviations were estimated significantly better by Response Process 3 and by the Consistent Response Process than by Response Process 1 and by Response Process 2.

The recovered stimulus standard deviations were estimated significantly better by Response Process 3 than any of the other response processes. However, the differences in the accuracy of the recovered stimulus standard deviations were small from a practical standpoint, *RMSD* = .336 to .356.

The average discrepancies of stimulus standard deviation estimates and category boundary standard deviation estimates can be compared because the true stimulus standard deviations and the true category boundary standard deviations had the same standard deviations. A similar pattern was seen for the standard deviations. Response process accounted for 64% and 10% of the within-family variation in estimation accuracy for category boundary standard deviations and stimulus standard deviations respectively, with a much larger amount of variability in estimation accuracy for category boundary standard deviations (SS_{wf} = 6.852) than for stimulus standard deviations (SS_{wf} = .267). Indeed there was approximately 26 times more variability in category boundary standard deviation estimation accuracy than in stimulus standard deviation estimation accuracy. These differences are shown for each response process in Figure 5.

The interaction between response process and number of stimuli had a small effect on the accuracy of the recovered parameters for stimulus standard deviations, η^2_{wf} = .101. Simple main effects were examined by first conditioning on response process and comparing accuracy differences between levels of number of stimuli. The familywise error rate was set at .0125. Using the Bonferroni correction, four independent-samples t-tests were conducted at α = (.0125 / 4) = .0031. Results indicate that stimulus standard deviations were estimated significantly better with 10 stimuli in the model than with 20 stimuli in the model for Response Process 2 and Response Process 3. There were no other statistically significant differences.

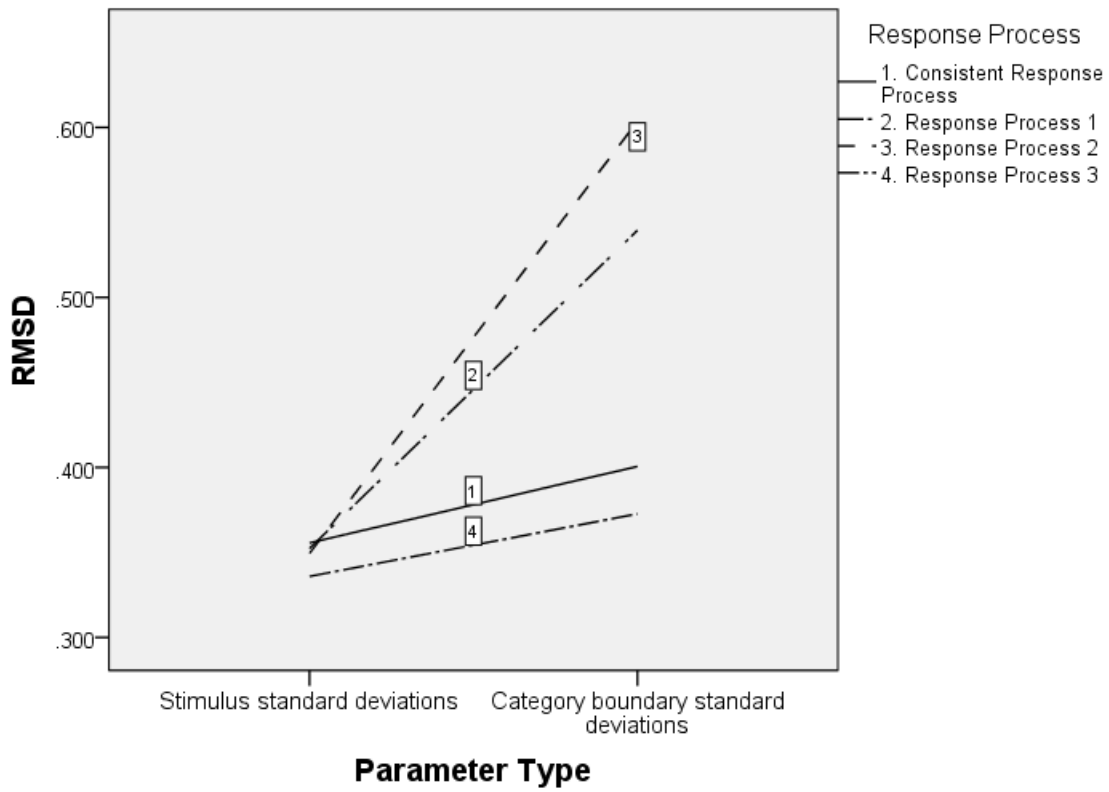


Figure 5. Root-mean-square deviations for stimulus standard deviations and category boundary standard deviations by response process.

Simple main effects for the response process by number of stimuli interaction were also examined by conditioning on the number of stimuli and comparing stimulus standard deviation accuracy differences between levels of response process. The familywise error rate was maintained at .0125. Using the Bonferroni correction, 12 paired-samples t-tests were conducted at $\alpha = (.0125 / 12) = .001$. Results indicate the stimulus standard deviations were estimated significantly better by Response Process 3 than by Response Process 1 and by the Consistent Response Process with 10 stimuli in the model. Further, stimulus standard deviations were estimated significantly better by Response Process 2 than by the Consistent Response Process with 10 stimuli in the model. Again, the actual differences in accuracy were small and the statistically

significant results may be due to the high power of the tests. A mean plot of the interaction is shown in Figure 6.

The number of stimuli had a large effect on the accuracy of the recovered category boundaries, η^2_{wf} = .253. The accuracy of the category boundary estimates decreased, RMSD = .359 to .492, as the number of stimuli in the model increased from 10 to 20. The number of response categories also had a large effect on the accuracy of the recovered category boundaries, η^2_{wf} = .314. The accuracy of the category boundary estimates increased, RMSD = .499 to .351, as the number of categories in the model increased from five to nine.

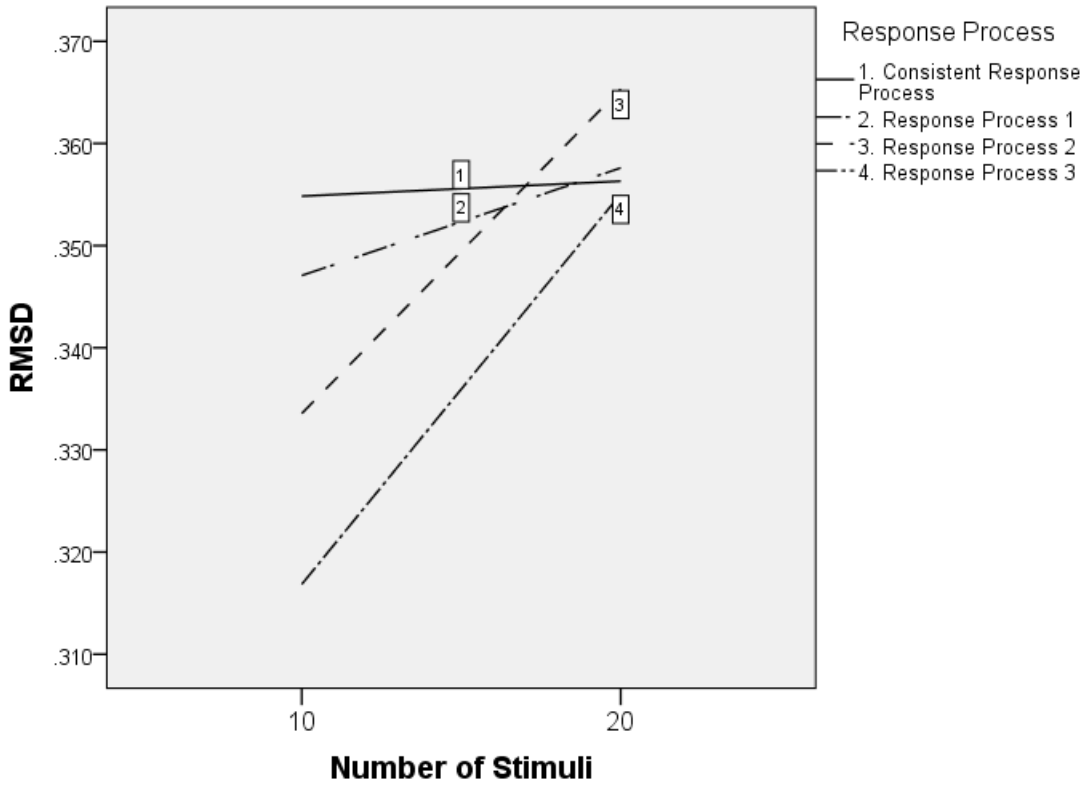


Figure 6. Root-mean-square deviations for 10 stimuli and 20 stimuli by response process.

Finally, the number of stimuli had a large effect on the accuracy of the recovered category boundary standard deviations, η^2_{wf} = .382. The accuracy of the category boundary standard deviation estimates decreased, RMSD = .392 to .567, as the number of stimuli in the model increased from 10 to 20.

Model Fit

Overall model fit was assessed using an internal consistency check given by Edwards and Thurstone (1952). First, a theoretical cumulative proportions matrix was calculated from the estimated parameters for each model. Then, the absolute discrepancies between the observed matrix and the theoretical matrix were calculated. The average absolute discrepancy served as an indicator for assessing overall model fit, with lower values indicating better fit of the model to the data. Essentially, the average absolute discrepancy measured the extent to which the observed response matrix could be accurately summarized by the parameter estimates.

Descriptive results indicate the model fit the data well when the Consistent Response Process was used to rate stimuli. Average absolute discrepancies were between .022 and .039, dependent on the number of stimuli and categories in the model. When Response Processes 1-3 were used to rate stimuli, the model fit the data reasonably well, but to a lesser extent than with the Consistent Response Process. Notably, models with nine categories fit the data better than models with five categories. The number of stimuli did not appear to affect the fit of the model to the data. The average absolute discrepancies are shown in Table 3.

Table 3

Average absolute discrepancy between theoretical and observed proportions for each model

	10S x 5C	**10S x 9C**	**20S x 5C**	**20S x 9C**
Consistent Response Process	0.033	0.022	0.039	0.022
Response Process 1	0.075	0.039	0.073	0.041
Response Process 2	0.094	0.038	0.086	0.040
Response Process 3	0.054	0.037	0.063	0.040

Note. S = number of stimuli; and C = number of response categories.

CHAPTER 4

DISCUSSION

The results from the parameter recovery study suggest that all parameters in the Law of Categorical Judgment (i.e. scale values, category boundaries, and the associated standard deviations) can be estimated accurately using the MCMC solution when the assumptions of Thurstone's scaling model are met. However, the assumptions of the model will likely not be met because of the discrepancy between the scaling model and the method by which stimulus judgments are obtained.

The parameter recovery study examined the effect of model violations on estimation accuracy by comparing estimates obtained from three response processes that violated the assumptions of the model with estimates obtained from a novel response process (i.e. The Consistent Response Process) that did not violate the assumptions of the model. Results suggest all parameters can be estimated reasonably well when these particular model violations occur, albeit to a lesser degree of accuracy than when the assumptions of the model are met. The following sections discuss the effects of response process, number of stimuli, and number of response categories on estimation accuracy.

The Effect of Response Process on Estimation Accuracy

The response process used to rate stimuli affected all four types of parameter estimates. However, there was much greater variability in the accuracy of the category boundary and category boundary standard deviation estimates than in the accuracy of the scale value and stimulus standard deviation estimates. Namely, there was approximately

15 times greater variability in the category boundary accuracies than in the scale value accuracies and approximately 26 times greater variability in the category boundary standard deviation accuracies than in the stimulus standard deviation accuracies.

These findings suggest that the discrepancy between the the method of successive intervals and Thurstone's scaling model does not affect the scale value and stimulus standard deviation accuracies in a practically meaningful way. However, the large amount of variability in the category boundary and category boundary standard deviation accuracies coupled with the large proportions of these variabilities explained by the response process used to rate the stimuli suggest that the discrepancy between the rating method and the scaling model assumptions does affect the category boundary and category boundary standard deviation estimates in a practically meaningful way.

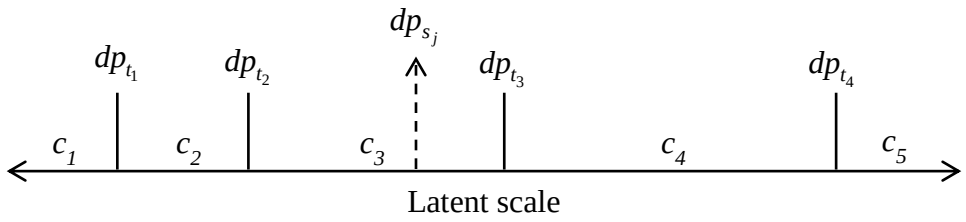
The large effect of response process on the category boundaries and category boundary standard deviations is not surprising considering the discrepancy between the rating method and the scaling model. According to the model, when a stimulus is rated, the discriminal process for the stimulus is compared with the discriminal process for each category boundary. If the stimulus is located between disordinal category boundaries, the rating depends on the response process utilized.

Consider the situation where the discriminal processes for category boundaries g and $g + 1$ are ordinal on a particular trial. Moreover, let the discriminal process for stimulus j fall between the two boundaries. The respondent rates the stimulus in category $g + 1$, which is the category directly above category boundary g and directly below category boundary $g + 1$. The explicit rating in the response frequency matrix (Table 1a) accurately represents the relationships between the stimulus and category boundary

discriminal processes. This leads to accurate implicit stimulus judgments, which are derived by summing the ratings in the response frequency matrix across rows from left to right and then dropping the last column of the matrix. Recall that when a respondent makes *J* explicit stimulus ratings according to the method of successive intervals, he or she provides *J*G* implicit stimulus judgments, each represented in the cumulative response frequency matrix as a "0" or a "1". A "0" indicates the stimulus is located above a given category boundary, whereas a "1" indicates the stimulus is located at or below a given category boundary. Each of the *J* rows corresponds with a stimulus, whereas each of the *G* columns corresponds with a category boundary after the last column of the cumulative response frequency matrix (Table 1b) is dropped.

Figure 7 shows how a given set of discriminal processes are coded at the individual level for a single stimulus when the stimulus falls between ordinal category boundaries. The locations of the discriminal processes associated with category boundaries are given by the vertical bars and the location of the discriminal process associated with stimulus *j* is given by the vertical arrow. In this case, the stimulus is located above the second boundary and below the third boundary. Therefore, according to the method of successive intervals, the individual rates stimulus *j* in response category 3. This response is coded with a one in the third column of the response frequency matrix and with zeros in the other columns of the matrix. The cumulative response frequency codes 0, 0, 1, 1, 1 are then derived by summing the codes in the response frequency matrix from left to right, and the implicit stimulus judgment codes are obtained by dropping the last column of the cumulative response frequency matrix, as shown in the diagram. The first two zeros in this string reflect the fact that the discriminal process for

Discriminal processes for stimulus *j* and category boundaries 1 to 4 when stimulus *j* is located between ordinal category boundaries *g = 2* and *g + 1 = 3*



Method of Successive Intervals

Response Frequency Matrix (row *j*)	Cumulative Response Frequency Matrix (row *j*)	Implicit Stimulus Judgments Matrix (row *j*)
c_1 c_2 c_3 c_4 c_5	c_1 c_2 c_3 c_4 c_5	t_1 t_2 t_3 t_4
s_j 0 0 1 0 0 →	0 0 1 1 1 →	0 0 1 1

Stimulus *j* rated in response category 3

Consistent Response Process

Explicit Stimulus Judgments Matrix (row *j*)

	t_1	t_2	t_3	t_4
s_j	0	0	1	1

Stimulus *j* judged to be at/below or above each category boundary

Figure 7. Derivation of implicit stimulus judgments from a single stimulus rating when stimulus *j* is located between ordinal category boundaries.
Note: s_j denotes stimulus *j*; t_g denotes category boundary *g*; c_g denotes response category *g*; dp_{s_j} denotes the discriminal process for stimulus *j*; and dp_{t_g} denotes the discriminal process for category boundary *g*.

the stimulus is above the first and second category boundaries, whereas the latter two ones in the string represent the fact that the discriminal process for the stimulus is at/below the third and fourth category boundaries.

Generally speaking, if the respondent explicitly rates stimulus *j* at or below category boundary *g*, the stimulus is also assumed to be located at or below all higher order category boundaries, $g + 1$ to *G*. Thus an explicit rating of stimulus *j* in category *g* of the response frequency matrix is represented in row *j* of the cumulative response frequency matrix as a series of zeros in columns 1 to $g - 1$ and ones in columns *g* to *G*. The cumulative response frequencies for the sample are obtained by summing all such response vectors for stimulus *j* across the *N* subjects.

When the successive intervals procedure is used, the implicit stimulus judgments illustrated in the diagram above are assumed based on the subject's categorization of the stimulus into a single response category. In contrast, when the Consistent Response Process is used to develop explicit stimulus judgments, the stimulus is explicitly compared to each category boundary. More importantly, when category boundaries are ordered, the implicit stimulus judgments implied by the successive intervals procedure and the explicit stimulus judgments obtained from the Consistent Response Process are identical.

Now consider the situation where the discriminal process for category boundary $g + 1 = 3$ falls below that for category boundary $g = 2$ on a particular trial. Moreover, let the discriminal process for stimulus *j* fall between these two boundaries. The respondent will either rate the stimulus in response category 2 because the category is located adjacent to and below category boundary 2 on the rating scale, or will rate the stimulus in response category 4 because the response category is located adjacent to and above category boundary 3 on the rating scale. The respondent will rate the stimulus in one of these response categories depending on the particular response process he or she uses to

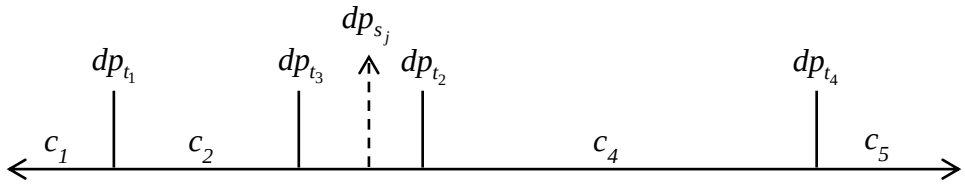
develop implicit stimulus judgments. However, regardless of the response category in which stimulus *j* is rated, the implicit stimulus judgments will not produce an implicit stimulus judgments matrix that accurately represents the relationships between the stimulus and category boundary discriminal processes on the latent scale.

Figure 8 shows how a given set of discriminal processes are coded at the individual level for a single stimulus when the stimulus falls between disordinal category boundaries $g + 1 = 3$ and $g = 2$. If the stimulus is rated in response category $g = 2$, the implicit stimulus judgment codes will be 0,1,1,1. The codes indicate that stimulus *j* is located above category boundary 1 and at/below category boundaries 2, 3, and 4. If stimulus *j* is instead rated in category $g + 2 = 4$, the implicit stimulus judgment codes will be 0,0,0,1. The codes indicate that the stimulus is located above category boundaries 1, 2, and 3, and at/below category boundary 4. However, neither of these implicit response code vectors accurately represent the relationships between stimulus *j* and the category boundaries on the latent scale. In reality stimulus *j* is located below category boundary $g = 2$ and above category boundary $g + 1 = 3$, so the stimulus judgment codes should be 0,1,0,1. These stimulus judgment codes can be obtained from the Consistent Response Process and accurately represent the relationships between stimulus *j* and the category boundaries on the latent scale.

As has been shown, when the discriminal process for stimulus *j* falls between disordinal category boundary discriminal processes and the stimulus is rated according to the successive intervals procedures, at least one of the momentary category boundary locations will be misrepresented in the implicit stimulus judgments matrix. Because the means and standard deviations of the category boundary distributions are estimated from

a transformation of the implicit stimulus judgments matrix, it is not surprising these parameters were misestimated in the parameter recovery study.

Discriminal processes for stimulus *j* and category boundaries 1 to 4 when stimulus *j* is located between disordinal category boundaries *g = 2* and *g + 1 = 3*



Method of Successive Intervals

Response Frequency Matrix (row *j*)	Cumulative Response Frequency Matrix (row *j*)	Implicit Stimulus Judgments Matrix (row *j*)
s_j | 0 1 0 0 0	→ 0 1 1 1 1	→ 0 1 1 1

Stimulus *j* rated in response category 2

Response Frequency Matrix (row *j*)	Cumulative Response Frequency Matrix (row *j*)	Implicit Stimulus Judgments Matrix (row *j*)
s_j | 0 0 0 1 0	→ 0 0 0 1 1	→ 0 0 0 1

Stimulus *j* rated in response category 4

Consistent Response Process

Explicit Stimulus Judgments Matrix (row *j*)

	t_1	t_2	t_3	t_4
s_j	0	1	0	1

Stimulus *j* judged to be at/below or above each category boundary

Figure 8. Derivation of implicit stimulus judgments from a single stimulus rating when stimulus *j* is located between disordinal category boundaries.
Note: s_j denotes stimulus *j*; t_g denotes category boundary *g*; c_g denotes response category *g*; dp_{s_j} denotes the discriminal process for stimulus *j*; and dp_{t_g} denotes the discriminal process for category boundary *g*.

The Consistent Response Process provided a means for examining the effect of innacurate information in the implicit stimulus judgments matrix on estimation accuracy because the *J*G* explicit stimulus judgments obtained from the Consistent Response Process accurately reflected the category boundary distributions on the latent scale. As would be expected from the explanation of implicit stimulus judgment matrix inaccuracies, the category boundaries and category boundary standard deviations obtained from the three response processes that violated the assumptions of the model were estimated to a lesser degree of accuracy than the estimates obtained from the Consistent Response Process. These results suggest that category boundary and category boundary standard deviations will be systematically misestimated to at least some degree when the method of successive intervals is used to obtain stimulus judgments to the extent that category boundary distributions overlap on the latent scale.

The Effect of Number of Stimuli on Estimation Accuracy

Curiously, the estimation accuracy of the category boundaries and category boundary standard deviations decreased as the number of stimuli in the model increased. This could be related to the inaccuracies in the implicit stimulus judgments matrix for Response Processes 1-3 because the number of stimuli in the model did not appear to affect the estimation accuracy of the category boundaries when the Consistent Response Process was used to obtain stimulus judgments. However, the inaccuracies in the implicit stimulus judgments matrix for Response Processes 1-3 does not explain why category boundary standard deviations were estimated less accurately for models with 20 stimuli than for models with 10 stimuli. Namely, less accurate category boundary standard

deviation estimates were obtained for the 20 stimuli models when the Consistent Response Process was used as well as when the inconsistent response processes were used to obtain stimulus judgments.

A future study could examine models with various numbers of stimuli (e.g., 10, 12, 14, 16, 18, & 20) to determine whether a stable relationship exists between the number of stimuli in the model and the estimation accuracy of the category boundaries and category boundary standard deviations.

The Effect of Number of Categories on Estimation Accuracy

Category boundaries were estimated more accurately with nine categories than with five categories in the model. A possible explanation for this finding is that the most extreme category boundaries (i.e. boundaries *1* & *G*) were poorly estimated for both the five category and nine category models, whereas the middle boundaries (i.e. boundaries *2* to *G – 1*) were estimated to a greater degree of accuracy. The nine category models had a larger number of middle boundaries than the five category models, so the misestimation of the extreme boundaries would have less of an effect on the overall estimation accuracy of the category boundaries for the nine category models than for the five category models.

The reason extreme category boundaries would be estimated less accurately than middle boundaries is because the category boundaries associated with the extreme columns of the z-score matrix (Table 1d) are estimated from less precise z-scores. The z-scores are less precise because the probit transformation used to convert the cumulative proportions matrix (Table 1c) to the z-score matrix is nonlinear and a .01 unit change in a

cumulative proportion results in a larger z-score change for low and high proportions than for moderate proportions. Because the first and *G*th columns of the cumulative proportions matrix, on average, have proportions closer to 0 and 1 respectively, the z-scores in the first and last columns of the z-score matrix will, on average, be less precise than the z-scores in the middle columns of the matrix.

A future study could examine whether extreme category boundaries are less accurately estimated than middle boundaries, and whether these differences in estimation accuracy can account for the overall differences in category boundary estimation accuracy observed between models with five and nine categories.

Corroborating Evidence from the Model Fit Assessment

The overall model fit assessment supported the primary findings of the current study. Namely, the observed cumulative proportions matrix was accurately recovered by the parameter estimates when the assumptions of the scaling model were met and was recovered to a lesser degree of accuracy when the assumptions of the scaling model were violated.

Although the number of stimuli in the model affected the estimation accuracy of the category boundaries and category boundary standard deviations, the number of stimuli did not appear to affect the recovery accuracy of the observed cumulative proportions matrix. The model fit results suggest the parameter estimating errors canceled out when reproducing the cumulative proportions matrix. Namely, the models with 20 stimuli recovered the observed cumulative proportions matrix to a similar degree of accuracy as the models with 10 stimuli.

The only parameters affected by the number of categories in the model were the category boundaries, which were estimated more accurately in the nine category models than the five category models. The number of categories also appeared to affect the recovery accuracy of the observed cumulative proportions matrix. Namely, the models with nine categories recovered the observed cumulative proportions matrix more accurately than the models with five categories. These findings are consistent with the notion that more information leads to more accurately estimated parameters, and this is subsequently translated into more accurate reproduction of the cumulative proportions.

Theoretical Implications

Benjamin, Diaz, and Wee (2009) gave evidence of the explanatory power gained from including category boundary variability in the signal detection theory model, a variant of Thurstone's scaling model. The current paper gives a Bayesian solution that estimates all parameters in the model, including category boundary standard deviations.

Recognizing the discrepancy between the scaling model and the rating method, Rosner and Kochanski (2009) gave a hypothetical response process that determined stimulus ratings when stimuli were located between disordinal category boundaries. Klauer and Kellen (2012) then pointed out that an infinite number of hypothetical response processes could be used to determine stimulus ratings when stimuli were located between disordinal category boundaries. They also emphasized that different response processes had different consequences for the predicted response distributions.

Although the hypothesized response processes help to explain the way respondents may rate stimuli that are located between disordinal category boundaries,

researchers have not examined the effect of these response processes on the estimation accuracy of the parameters. The current study used a novel method called the Consistent Response Process to obtain stimulus judgments that are consistent with the assumptions of Thurstone's scaling model. Conceivably, the method can be used in experimental situations, albeit at a cost. Namely, the Consistent Response Process requires *J*G* stimulus judgments, whereas the method of successive intervals requires only *J* stimulus judgments. However, the Consistent Response Process is still more efficient than the method of paired comparisons, which requires *J*(J – 1)/2* stimulus judgments, when the number of stimuli in the model is greater than or equal to 10.

The Consistent Response Process requires a large number of stimulus judgments because the respondent compares each stimulus to each category boundary. Consider a situation where attitude statements are scaled with respect to their favorability toward an attitude object. Suppose that favorability judgments were obtained with five graded categories where 1 = Very Unfavorable, 2 = Unfavorable, 3 = Neutral, 4 = Favorable, and 5 = Very Favorable. For a particular attitude statement, the respondent judges whether the statement is better described as "Very Unfavorable" or "Unfavorable" toward the attitude object. Note that these two categories are located below and above the first category boundary respectively. Endorsing the category "Very Unfavorable" indicates that the attitude statement is located below the first category boundary on the latent scale, whereas endorsing the category "Unfavorable" indicates that the attitude statement is located above the first category boundary. After making the first stimulus judgment, the respondent then judges whether the attitude statement is better described as "Unfavorable" or "Neutral" toward the attitude object. This method continues until the

respondent finally judges whether the attitude statement is better described as "Favorable" or "Very Favorable" toward the attitude object. Thus the respondent makes *G* total explicit judgments about the location of stimulus *j* on the latent scale in relation to each of the *G* category boundaries. In practice, the same stimulus would not be presented for trials *n* and *n + 1*, because this would increase dependency between the stimulus judgments.

Although the discriminal process for stimulus *j* cannot be directly compared to each of the *G* category boundary discriminal processes on a single trial in an experimental situation, the discriminal process for stimulus *j* can be directly compared to each of the *G* category boundary discriminal processes across *G* trials. Because the judgment associated with stimulus *j* and category boundary *g + 1* does not depend on the judgment associated with stimulus *j* and category boundary *g*, the category boundary distributions on the latent scale will not be misrepresented in the explicit stimulus judgments matrix. Further, the total number of discriminal processes in the experimental situation will be higher than the total number of discriminal processes in the parameter recovery study because discriminal processes for stimuli and category boundaries were only drawn once for each simulated subject in the parameter recovery study. The increase in the total number of discriminal processes in the experimental situation will likely lead to more accurate estimation of the means and standard deviations of the stimulus and category boundary distributions because more data will be available to calculate these statistics.

In the current study, estimates obtained from various response processes that are inconsistent with the scaling model were compared to estimates obtained from the

Consistent Response Process. The results indicate that parameters are estimated to a lesser degree of accuracy when inconsistent response processes are used to rate stimuli, and estimated particularly less accurately for category boundaries and category boundary standard deviations. Misestimation varied across the three inconsistent response processes, suggesting the particular response process used by a subject has different consequences on the parameter estimates.

Applied Implications

Although it is unlikely that participants are actually aware of disordinal category boundaries on a given trial, participants likely experience rating ambiguity as a result of category boundary variability. The only conditions in which disordinal category boundaries (and the associated rating ambiguity) would not occur under the Thurstone model are when category boundary distributions do not overlap and/or category boundary distributions have constant variance and are perfectly correlated. As these conditions seem unlikely to be met in psychological or psychophysical tasks, it seems reasonable to assume that category boundaries will be located close enough on a given trial to prompt rating ambiguity from the respondent.

Consider the situation where the respondent is uncertain about the location of stimulus j in relation to disordinal category boundaries g and $g + 1$ on a given trial. Moreover, let the discriminal process for stimulus j be located between those for these disordinal category boundaries. The respondent would rate the stimulus in one of two categories (i.e. category g or category $g + 2$) depending on the particular response process he or she uses. As discussed previously, regardless of the category in which the

stimulus is rated, the implicit stimulus judgments in the implicit stimulus judgments matrix will not accurately represent the relationships between the stimulus and category boundary discriminal processes on the latent scale. Moreover, these types of misrepresentations are, in reality, violations of the distributional assumptions in the Thurstone model.

The parameter recovery study showed that the aforementioned model violations can lead to misestimation of the model parameters. In practice, researchers are primarily concerned with scale values and stimulus standard deviations because these parameters can be used to measure individual differences in the Thurstone paradigm. Fortunately, the scale values and stimulus standard deviations do not appear to be affected by model violations to the same extent as the category boundaries and category boundary standard deviations. However, researchers interested in understanding the way participants perceive the rating scale or how stimuli are perceived in relation to the scale may benefit from using the Consistent Response Process to obtain stimulus judgments. If the cost of using the Consistent Response Process is too high, this study will at least serve as a cautionary note that estimation inaccuracies can occur when category boundaries vary across trials and the method of successive intervals is used to obtain stimulus judgments.

Limitations of the Current Study and Future Directions

The current study examined a theoretical problem that could only be investigated through a parameter recovery study. To examine the discrepancy between the assumptions of Thurstone's scaling model and the method by which stimulus judgments are obtained, data were generated according to the assumptions of Thurstone's model.

Namely, stimuli and category boundaries were assumed to be normally and independently distributed. In practice, stimuli and category boundaries may not be normally distributed and are likely to covary to some extent.

Thurstone's scaling model would more accurately reflect the way stimuli are perceived in relation to the rating scale if the stimuli and category boundaries were allowed to correlate. However, estimating a correlation matrix may not be feasible because the number of unknown parameters in a model with correlated stimuli and category boundaries exceeds the number of known normal deviates in the z-score matrix. The following inequality, which compares the number of unknown parameters in the LCatJ on the left side and the number of known normal deviates on the right side, will hold regardless of the number of stimuli and category boundaries in the model.

$$2(J+G)-2+\frac{(J+G)(J+G-1)}{2} > JG \tag{8}$$

where

J is the number of stimuli; and

G is the number of category boundaries.

However, the number of unknown parameters can be less than or equal to the number of known normal deviates if some of the correlations are constrained, in which case, the parameters may be estimable. For example, the number of unknown parameters in the model can be reduced by constraining the correlations between stimuli, as well as the correlations between stimuli and category boundaries, and estimating only the correlations between category boundaries in addition to the other parameters estimated in the current study. This leads to the following inequality:

$$2(J+G)-2+\frac{(G)(G-1)}{2}\leq JG. \quad (9)$$

The inequality is true when there are at least three category boundaries and seven stimuli in the model. The minimum number of stimuli required for the inequality to be true varies as a function of the number of category boundaries in the model. For example, with three category boundaries, seven stimuli are required for the inequality to be true, but for four category boundaries, only six stimuli are required for the inequality to be true. A scatterplot of the relationship between the number of category boundaries and the minimum number of stimuli required for the inequality to be true is shown in Appendix C.

Estimating the correlations between category boundaries would provide information about the extent to which rating ambiguity occurs as a function of disordinal category boundaries. If the category boundary distributions overlap, strong correlations between category boundaries would suggest that on a given trial, there is a low probability of disordinal category boundaries, whereas weak correlations between category boundaries would suggest that on a given trial, there is a higher probability of disordinal category boundaries. Further, if the category boundaries are highly correlated, there will be fewer misrepresentations of the category boundary distributions in the cumulative response frequency matrix, and subsequently, the category boundaries and category boundary standard deviations will be misestimated to a lesser extent.

The model may also be improved through the implementation of different constraints for setting the origin and unit of the latent scale. In the current study, relatively loose constraints were set, affecting only the first category boundary and the first category boundary standard deviation (i.e. the parameters were fixed to zero and

unity respectively). It would be worthwhile to examine the implementation of tighter constraints, such as fixing the mean of all scale values and category boundaries to zero and the mean of all stimulus and category boundary variances to unity. Although difficult to implement in OpenBUGS, these constraints may better keep the estimates from drifting to unrealistic values when insufficient information is provided by the z-score matrix.

Finally, further research on this topic should investigate the utility of the Bayesian solution to the LCatJ given in the current paper. When real category boundary variability exists, the MCMC solution may estimate scale values and standard deviations more accurately than previous solutions. A parameter recovery study comparing the MCMC solution given in the current paper to previous solutions that do not account for category boundary variability would shed light on whether the scale values and stimulus standard deviations are estimated more accurately when category boundary variability is accounted for in the model.

APPENDIX A

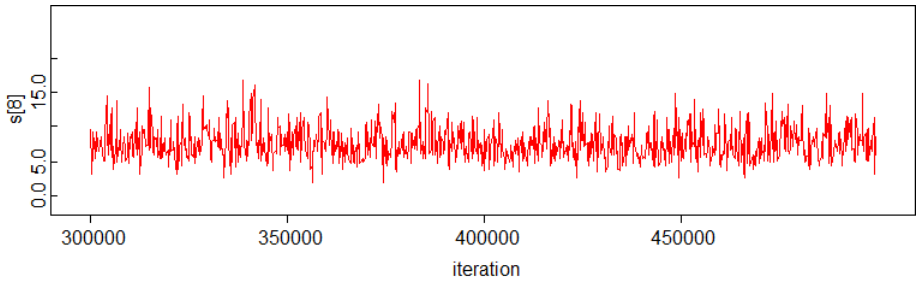
DIAGNOSTICS FOR ASSESSING THE CONVERGENCE OF THE MARKOV CHAIN

Table 4

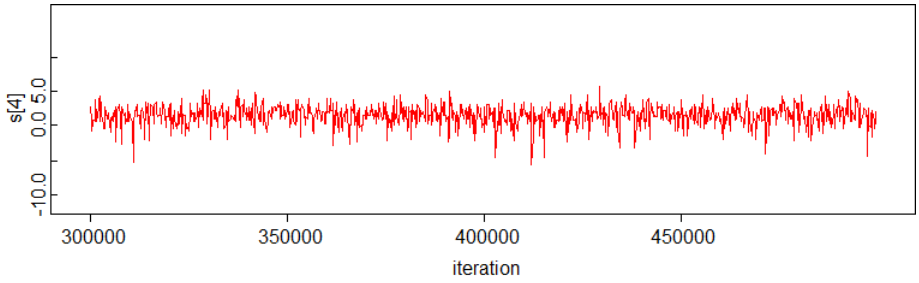
Proportion of statistical convergence tests passed for each stimulus by category condition

	Geweke's	Heidelberger and Welch's	
	z-test	Stationarity test	Half-width test
10 stimuli, 5 categories	0.92	1.00	0.81
10 stimuli, 9 categories	0.82	1.00	0.91
20 stimuli, 5 categories	0.89	1.00	0.89
20 stimuli, 9 categories	0.85	1.00	0.91

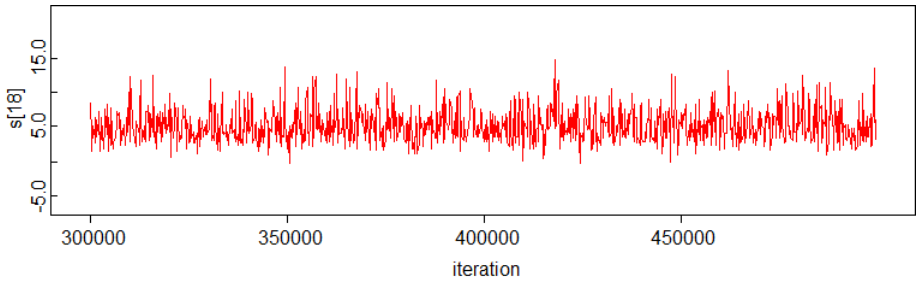
(a)



(b)



(c)



(d)

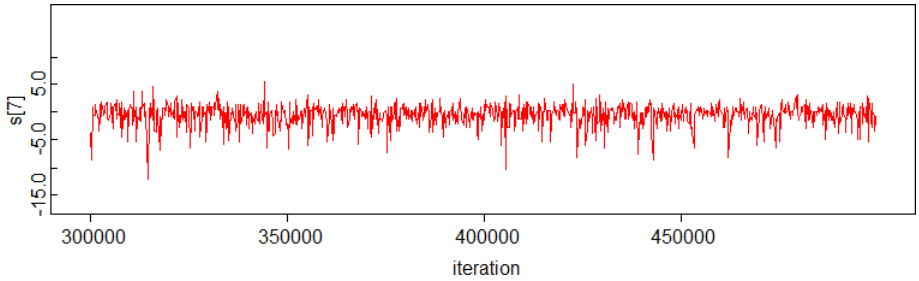
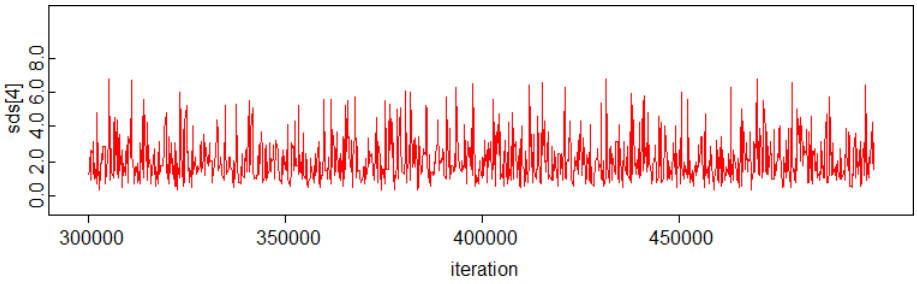
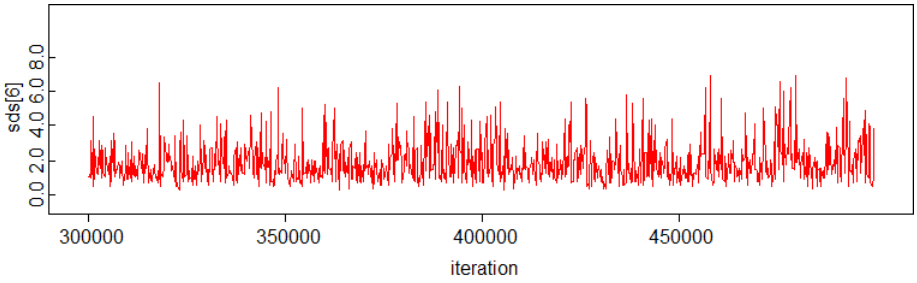


Figure 9. Prototypical trace plots for scale value posterior distributions, including the (a) best and (b) worst plots for the 10 stimuli by 9 category boundaries condition and the (c) best and (d) worst plots for the 20 stimuli by 5 category boundaries condition.

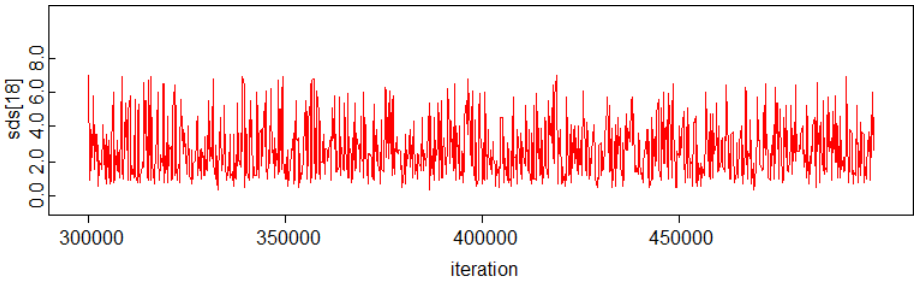
(a)



(b)



(c)



(d)

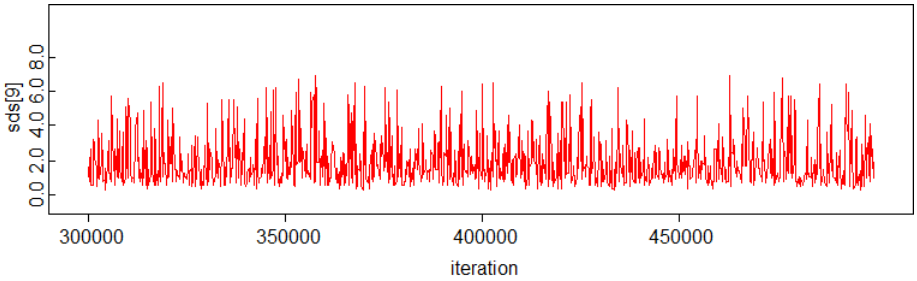
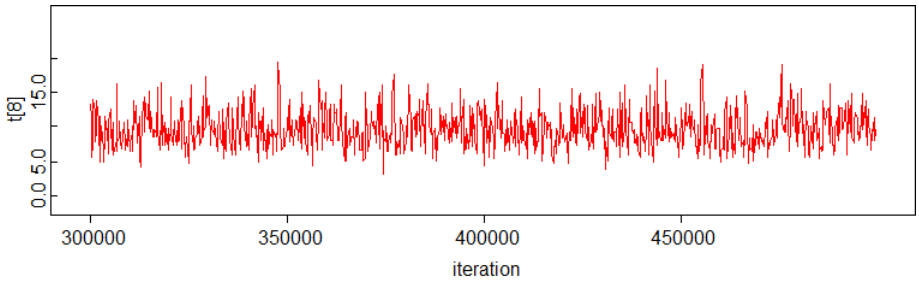
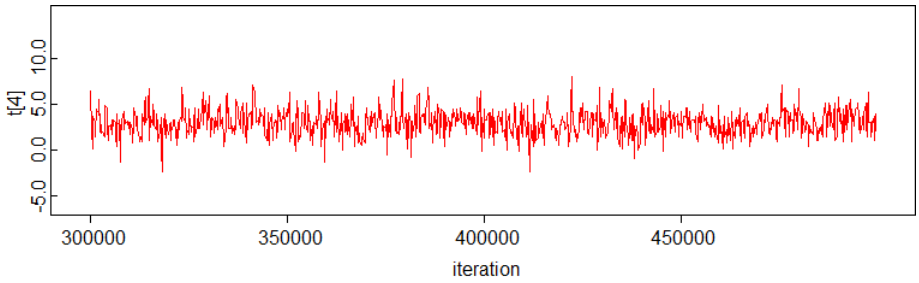


Figure 10. Prototypical trace plots for stimulus standard deviation posterior distributions, including the (a) best and (b) worst plots for the 10 stimuli by 9 category boundaries condition and the (c) best and (d) worst plots for the 20 stimuli by 5 category boundaries condition.

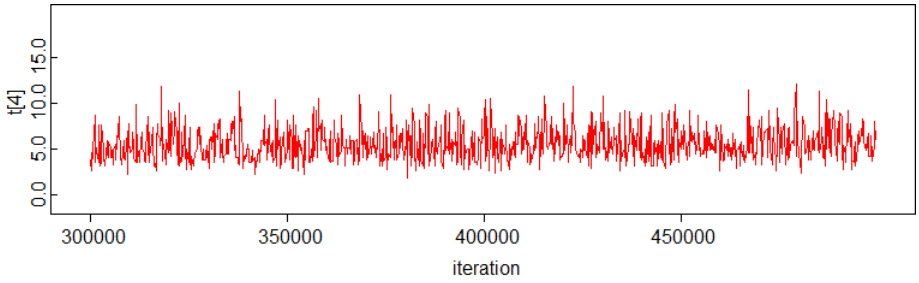
(a)



(b)



(c)



(d)

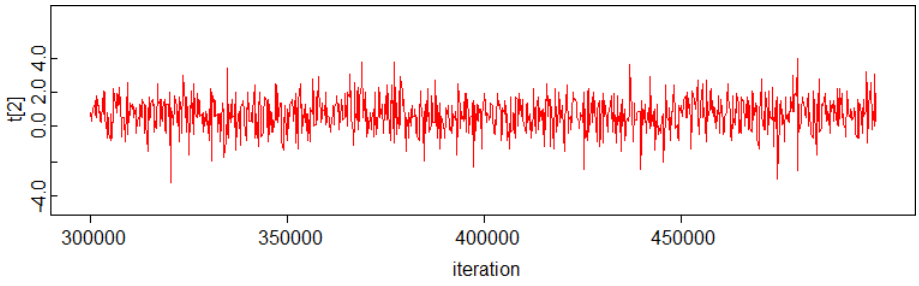
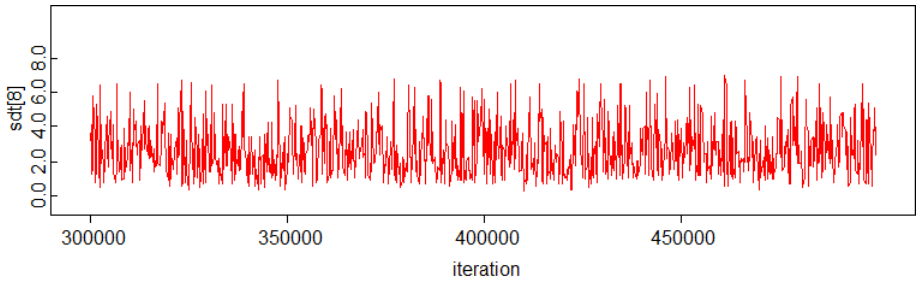
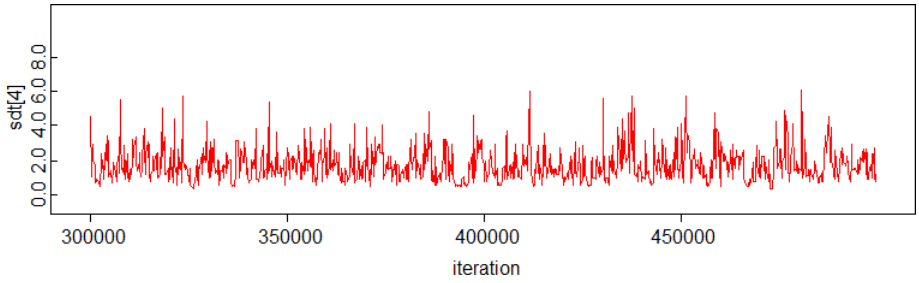


Figure 11. Prototypical trace plots for category boundary posterior distributions, including the (a) best and (b) worst plots for the 10 stimuli by 9 category boundaries condition and the (c) best and (d) worst plots for the 20 stimuli by 5 category boundaries condition.

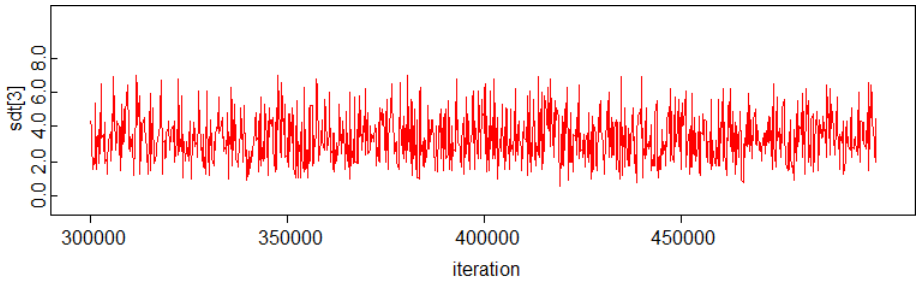
(a)



(b)



(c)



(d)

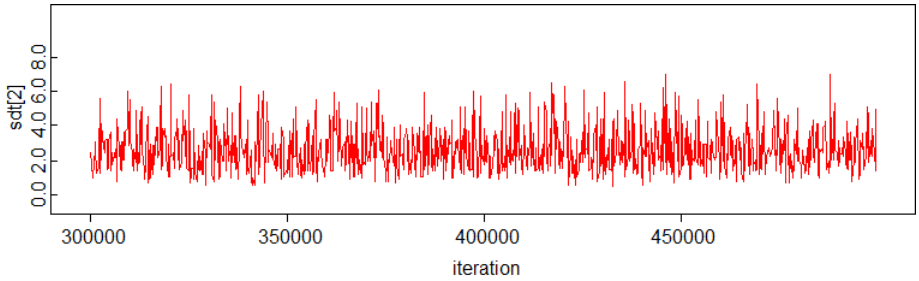


Figure 12. Prototypical trace plots for category boundary standard deviation posterior distributions, including the (a) best and (b) worst plots for the 10 stimuli by 9 category boundaries condition and the (c) best and (d) worst plots for the 20 stimuli by 5 category boundaries condition.

APPENDIX B

MEAN ACCURACY MEASURES FOR ESTIMATED PARAMETERS

Table 5

Mean accuracy measures for estimated scale values and category boundaries by the main effects of response process, number of stimuli, and number of categories

	RMSD		Correlation		Absolute Mean Diff.		Variance Ratio	
	$\hat{s}_j$	$\hat{t}_g$	$\hat{s}_j$	$\hat{t}_g$	$\hat{s}_j$	$\hat{t}_g$	$\hat{s}_j$	$\hat{t}_g$
RP								
Consistent	.216	.152	.993	.998	.018	.052	.977	1.144
One	.268	.494	.992	.993	.046	.108	.874	2.165
Two	.332	.689	.993	.989	.062	.236	.770	2.703
Three	.248	.365	.993	.997	.100	.305	.929	1.230
Stimuli								
Ten	.261	.359	.994	.995	.068	.146	.890	1.604
Twenty	.271	.492	.992	.993	.046	.204	.885	2.018
Categories								
Five	.271	.499	.989	.994	.046	.217	.845	2.428
Nine	.261	.351	.997	.994	.068	.134	.930	1.193

Note. RMSD = root-mean-square deviation; $\hat{s}_j$ = scale value estimates; $\hat{t}_g$ = category boundary estimates; RP = response process; stimuli = number of stimuli, and categories = number of categories.

Table 6

Mean accuracy measures for estimated stimulus and category boundary standard deviations by the main effects of response process, number of stimuli, and number of categories

	RMSD		Correlation		Absolute Mean Diff.		Variance Ratio	
	$\hat{\sigma}_j$	$\hat{\sigma}_g$	$\hat{\sigma}_j$	$\hat{\sigma}_g$	$\hat{\sigma}_j$	$\hat{\sigma}_g$	$\hat{\sigma}_j$	$\hat{\sigma}_g$
RP								
Consistent	.356	.401	.685	.815	.117	.332	.225	.626
One	.352	.539	.641	.389	.109	.405	.233	.299
Two	.349	.604	.695	.106	.111	.441	.188	.682
Three	.336	.373	.712	.485	.109	.193	.251	.381
Stimuli								
Ten	.338	.392	.694	.509	.130	.230	.271	.418
Twenty	.359	.567	.673	.388	.092	.455	.177	.576
Categories								
Five	.365	.509	.679	.459	.094	.417	.071	.554
Nine	.332	.449	.687	.439	.129	.269	.377	.440

Note. RMSD = root-mean-square deviation; $\hat{\sigma}_j$ = stimulus standard deviation estimates; $\hat{\sigma}_g$ = category boundary standard deviation estimates; RP = response process; stimuli = number of stimuli, and categories = number of categories.

Table 7

Mean accuracy measures for estimated scale values and category boundaries by the interaction effect of number of stimuli by response process

	RMSD		Correlation		Absolute Mean Diff.		Variance Ratio	
	$\hat{s}_j$	$\hat{t}_g$	$\hat{s}_j$	$\hat{t}_g$	$\hat{s}_j$	$\hat{t}_g$	$\hat{s}_j$	$\hat{t}_g$
Ten Stimuli								
CRP	.202	.132	.994	.998	.024	.052	.990	1.049
RP1	.264	.423	.994	.994	.065	.116	.873	1.904
RP2	.339	.579	.995	.991	.066	.174	.759	2.339
RP3	.240	.302	.995	.996	.115	.243	.938	1.123
Twenty Stimuli								
CRP	.230	.173	.992	.998	.011	.052	.965	1.240
RP1	.272	.566	.991	.991	.026	.100	.875	2.426
RP2	.325	.799	.992	.987	.059	.299	.781	3.068
RP3	.256	.429	.992	.997	.085	.367	.919	1.337

Note. RMSD = root-mean-square deviation; $\hat{s}_j$ = scale value estimates; $\hat{t}_g$ = category boundary estimates; CRP = Consistent Response Process; and RP1-3 = Response Processes 1-3.

Table 8

Mean accuracy measures for estimated stimulus and category boundary standard deviations by the interaction effect of number of stimuli by response process

	RMSD		Correlation		Absolute Mean Diff.		Variance Ratio	
	$\hat{\sigma}_j$	$\hat{\sigma}_g$	$\hat{\sigma}_j$	$\hat{\sigma}_g$	$\hat{\sigma}_j$	$\hat{\sigma}_g$	$\hat{\sigma}_j$	$\hat{\sigma}_g$
Ten Stimuli								
CRP	.355	.309	.698	.823	.157	.217	.273	.606
RP1	.347	.431	.644	.437	.140	.263	.288	.218
RP2	.334	.485	.698	.215	.112	.299	.232	.538
RP3	.317	.342	.735	.561	.113	.143	.293	.310
Twenty Stimuli								
CRP	.356	.492	.672	.808	.076	.448	.177	.646
RP1	.358	.648	.638	.340	.077	.547	.178	.379
RP2	.365	.722	.692	-.004	.110	.583	.144	.826
RP3	.355	.404	.690	.409	.105	.242	.208	.452

Note. RMSD = root-mean-square deviation; $\hat{\sigma}_j$ = stimulus standard deviation estimates; $\hat{\sigma}_g$ = category boundary standard deviation estimates; CRP = Consistent Response Process; and RP1-3 = Response Processes 1-3.

APPENDIX C

MINIMUM PARAMETER REQUIREMENTS FOR ESTIMATING A RESTRICTED CORRELATION MATRIX

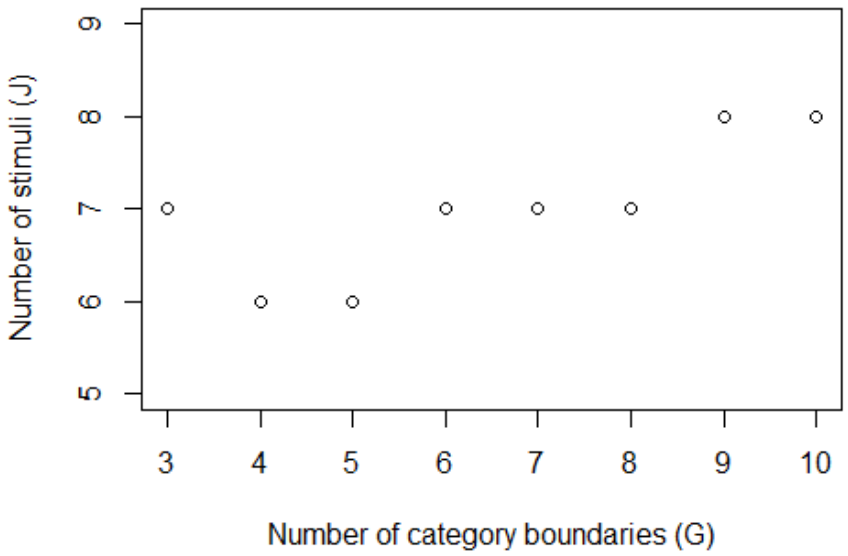


Figure 13. Minimum number of stimuli required for the number of known normal deviates to equal or exceed the number of unknown parameters when a correlation matrix is estimated for category boundaries.

APPENDIX D

OPENBUGS CODE AND EXAMPLE DATA

```
model
{
# specifying the relationship between the data and the model parameters
for (j in 1:J)
        {
        for (k in 1:K)
                {
                z[j,k] <- (  t[k] - s[j] ) / (sqrt((1/ps[j]) + (1/pt[k])));
                x[j,k] ~ dnorm(z[j,k],1);
                }
        }
# specifying prior distributions for model parameters
for (j in 1:J)
        {
        s[j] ~ dnorm (.01,.01);
        ps[j] ~ dgamma(.5,.5)I(0.0204081633,);
        sds[j] <- sqrt(1/ps[j]);
        }
# specifying the origin and unit of the scale by constraining the first category
# boundary to zero and the first category boundary standard deviation to
# unity
t[1] <- 0;
pt[1] <- 1;
sdt[1] <- sqrt(1/pt[1]);
for (k in 2:K)
        {
        t[k] ~ dnorm(.01,.01);
        pt[k] ~ dgamma(.5,.5)I(0.0204081633,);
        sdt[k] <- sqrt(1/pt[k]);
        }
}
```

Example data for model with 10 stimuli and 5 response categories

```
list(x = structure(
        .Data = c(
        0.7789655,   NA      ,   NA      ,   NA      ,
        0.1763741,   0.8203791,  1.1952227,  1.1749867,
        -0.1509692,  0.6495235,  1.0984684,  1.1952227,
        -0.3002322,  0.2326927,  0.6188730,  0.9077695,
        -0.9384756,  -0.1357739, 0.5301614,  0.7257370,
        -1.0194276,  -0.4288945, 0.1307159,  0.5129304,
        -1.4466320,  -0.8559959, -0.2378466, 0.1004337,
        NA      ,    -1.4610562, -0.7789655, -0.1105162,
        NA      ,    NA      ,   -1.2702376, -0.4454425,
        NA      ,    NA      ,   -1.4050715, -0.6007597
        ),
        .Dim = c(10 , 4 )), J = 10 , K = 4 )
```

REFERENCES

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*, 84-115.

Bock, R. D., & Jones, L. V. (1968). The method of successive categories. In *The measurement and prediction of judgment and choice* (pp. 212-247). San Francisco: Holden Day, Inc.

Bockenholt, U. (2006). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika, 71*, 615-629.

Edwards, A. L., & Thurstone, L. L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17*, 169-180.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. Smith (Eds.), *Bayesian Statistics 4*. Oxford, UK: Clarendon Press.

Gill, J. (2002). *Bayesian methods. A social and behavioral sciences approach*. Boca Raton : Chapman & Hall/CRC.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31*, 1109-44.

Jones, L. V. (1959). Some invariant findings under the method of successive intervals. *The American Journal of Psychology, 72*, 210-220.

Klauer, K. C., & Kellen, D. (2012). The Law of Categorical Judgment (corrected) extended: A note on Rosner and Kochanski (2009). *Psychological Review, 119*, 216-220.

Kochanski, G., & Rosner, B. S. (2010). Bootstrap Markov chain Monte Carlo and optimal solutions for the Law of Categorical Judgment (Corrected). Retrieved from http://arxiv.org/abs/1008.1596.

Lee, W. (1969). Relationships between Thurstone category scaling and Signal Detection Theory. *Psychological Bulletin, 71*, 101-107.

Mackay, D. B., & Chaiy, S. (1982). Parameter estimation for the Thurstone Case III model. *Psychometrika, 47,* 353-359.

McFadden, D. (2001). Economic choices. *The American Economic Review, 91*, 351-378.

McNicol, D. (1972). *A primer of signal detection theory*. London: George Allen & Unwin.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News, 6,* 7-11.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–367.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied PsychologicalMeasurement, 20,* 231-255.

Rosner, B. S., & Kochanski, G. (2009). The Law of Categorical Judgment (corrected) and the interpretation of changes in psychophysical performance. *Psychological Review, 116,* 116-128.

Saffir, M. A. (1937). Comparative study of scales by three psychophysical methods. *Psychometrika, 2,* 179-198.

Sjoberg, L. (1964). Successive intervals scaling with unequal variances and covariances. *Scandinavian Journal of Psychology, 5,* 530-538.

Tanner, W. P. (1956). Theory of Recognition. *Journal of the Acoustical Society of America, 28,* 882-888.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review, 34,* 273-286.

Thurstone, L. L. (1927b). Psychophysical analysis. *The American Journal of Psychology, 38,* 368-389.

Torgerson, W. S. (1958). The law of categorical judgment. In *Theory and methods of scaling,* (pp. 205-46). New York, New York: John Wiley and Sons.

Yao, G., & Bockenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology, 52,* 79-92.